Exploring In-Service Science Teachers' Assessment Literacy on Teaching with Models

by

Alexis Miguel Gonzalez Donoso

M Ed., University of Santiago, Chile, 2012

B. Sc., University of Santiago, Chile, 2009

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Curriculum Studies)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

December 2021

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Exploring In-Service Science Teachers' Assessment Literacy on Teaching with Models

submitted by Alexis Miguel Gonzalez Donoso in partial fulfillment of the requirements for

the degree of Doctor of Philosophy

in Curriculum Studies

Examining Committee:

Dr. Samia Khan, Associate professor, Department of Curriculum and Pedagogy, UBC
Supervisor

Dr. Jackie Stewart, Associate professor, Department of Chemistry, UBC
Supervisory Committee Member

Dr. Yan Liu, Assistant professor, Department of Psychology, Carleton University
Supervisory Committee Member

Dr. Ann Anderson, Associate professor, Department of Curriculum and Pedagogy, UBC
University Examiner

Dr. Gregory Dake, Associate professor, Department of Chemistry, UBC
University Examiner

Abstract

Model-based Teaching (MBT) is an approach to teaching science that promotes the generation, evaluation, and modification of students' mental models (GEM cycle). This study is a sequential mixed-methods study to explore in-service science teachers (ISTs)' assessment literacy about teaching with models. A questionnaire was administered to 416 ISTs from Chile and Canada to investigate their knowledge of models in science education and how often they assessed their students' reasoning with models. Then, a focused investigation on the development of five Chilean ISTs' assessment literacy in MBT was undertaken. This investigation involved classroom observations and interviews before and after participating in an online professional development course on MBT. Quantitative data from the questionnaires were analyzed using regression analysis and factor analysis. A cross-case analysis was conducted with the five teachers to compare their assessment literacy. Results of the questionnaire showed that ISTs' knowledge of models and modeling was positively and significantly related to their assessment literacy. Regarding the five ISTs' pedagogy, it was found that most ISTs had beginner levels of proficiency in assessment literacy in MBT. After attending the online course, they continued using models to convey information. One experienced teacher; however, promoted a full GEM cycle which indicated that her enactments might have been influenced by her pedagogical content knowledge and years of teaching experience. These results showed that when ISTs are more literate in MBT, it influences their pedagogy regarding how they promote student generation and evaluation of their own models. Specifically, evidence was found that ISTs use assessment to i) judge students' reasoning with models; ii) communicate feedback to clarify students' conceptual doubts; iii) give opportunities to express their models and iv) promote the revision of generated models, for example, through the evaluation of models' predictive power

and revising them to fit new evidence. This study is significant in science education because it offers a new validated instrument to characterize ISTs' assessment literacy. Furthermore, the characterization of ISTs' assessment literacy offers an opportunity to identify which aspects of ISTs' MBT could benefit from further enrichment through science teacher education.

Lay Summary

I investigated in-service science teachers (ISTs)' assessment practices during their model-based teaching. A questionnaire was administered to Chilean and Canadian ISTs. Statistical analysis showed that when ISTs know more about models, they more often engage in assessment practices of students' models. Furthermore, I observed ISTs' instruction before and after attending a 10-hour online course in model-based teaching. The quantitative and qualitative findings revealed that ISTs mostly used assessment to judge students' understanding of models taught in class. Nevertheless, a case of an experienced IST showed that after the course, she engaged students in modeling. Her case showed that her knowledge about how to teach with models and years of teaching experience influenced the type of strategies used to assess and engage students in the generation, evaluation, and modification of models.

Preface

This dissertation is original, and the design, data collection and analysis are work by Alexis Miguel González Donoso. No chapters of this dissertation have been published in articles. This study was approved by the Behavioral Research Ethics Board at the University of British Columbia on May 27, 2019. Approval number: H18-02512.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to cordially express my appreciation to my supervisor Dr. Samia Khan for her guidance and support in this dissertation and her theoretical contribution in the framework used to conduct this study. I would also like to thank my committee member Dr. Yan Liu and Jackie Stewart for their valuable contribution to the methodology of the study. In addition, I would like to thank Drs. Ann Anderson, Gregory Dake and Ainoa Marzábal for their insightful comments to enrich the dissertation.

I would like to express my appreciation to the in-service science teachers in Canada and Chile who participated in the study. I am particularly grateful to those five in-service science teachers who allowed me to observe their pedagogy. Also, I gratefully acknowledge the administrative support of Alan Jay, graduate program assistant.

This work was funded by the National Agency for Research and Development (ANID) / Scholarship Program / DOCTORADO BECAS CHILE/2016 – 72170096.

Dedication

I dedicate this dissertation to my family, specially to my parents, Zoila Donoso and Eugenio González, for their support and inspiration.

Chapter 1: Introduction of the Study

1.1 Introduction

Teachers' knowledge about how to elicit and characterize students' models is of pivotal importance in helping students generate models that are useful for comprehending the real world. Model-based teaching or MBT, engages students in the development of epistemic practices in science, such as modeling, through the expression of their mental models (Johnson-Laird, 1980; Justi & Gilbert, 2002a; Khan, 2007; Windschitl, Thompson & Braaten, 2008). Khan (2011a) defines MBT as a theoretical approach to teaching which includes teachers' actions to promote students' generation, evaluation and critique of their mental models or cognitive representations of systems. In science education, the concept of model has two different meanings. This word can be used as a noun (Lattery, 2017) when attempting to show a representation of an object, mechanism, process, or a system (i.e., when students study the model of the Earth). Also, it can be used as a verb when involving the action of developing or displaying a model (i.e., when students are asked to model the structure of an atom). Models are representations that depict a set of assumptions and relationships (Khan, 2007) and are conceived as a component of scientific theories that are rich in scientific knowledge and help people think scientifically, solve problems, and communicate ideas by enriching their modeling practices through the expression, revision, and modifications of their models (Oh & Oh, 2011). This research aims to contribute to science education by analyzing in-service science teacher or ISTs' literacy in terms of assessing students' reasoning with models.

The concept of assessment literacy was suggested by Stiggins (1991b) who referred it as teacher capacity to understand whether assessment informs students about the achievement outcomes

that teachers' value, and a teacher's capacity to use the results from assessment to inform his/her/their practice. In this study, I used the framework of assessment literacy from Xu and Brown (2016) wherein assessment literacy is comprised of *i) disciplinary knowledge and Pedagogical Content Knowledge (PCK), ii) knowledge of assessment purposes, content and methods, iii) knowledge of grading, iv) knowledge of feedback, v) knowledge of peer and self-assessment, vi) knowledge of assessment interpretation and communication, and vii) knowledge of assessment ethics.*

The overarching objective of the study was to examine how working or in-service science teachers' (ISTs') assessment literacy is influenced by their understanding of models and in what ways does this assessment literacy influence their pedagogy to engage students in modeling. The research questions that guide this study are dual-fold.

> **Research Question 1:** Are ISTs' knowledge of models and modeling related to ISTs' assessment literacy in MBT?
>
> **Research Question 2:** In what ways do ISTs' assessment literacy about models and modeling influence their pedagogy?

A study of ISTs' assessment literacy can help us to establish a relationship between teaching and learning and teachers' pedagogy in science because this type of knowledge (being literate about assessment) allows teachers to identify the level of learning that students have achieved and help them make decisions about the learning intentions and course of action they must take to improve students' learning (Black & Wiliam, 2018). To identify teachers' assessment literacy in MBT, otherwise referred to in this dissertation as "ALMBT", the research design of this study

includes two phases. Firstly, a baseline phase was undertaken in order to gauge science teachers' knowledge about assessing models in the science classroom. Secondly, teachers' pedagogy was observed to provide evidence of how ISTs apply their knowledge of assessment while modeling in the science classroom. In the study participated Canadian and Chilean teachers to establish a baseline of teacher's knowledge of assessment through a questionnaire and then focused the study on Chilean teachers for classroom observation of their pedagogy. The second phase of the study continued with Chilean ISTs to characterize their pedagogy and identify how their ideas about assessment of models were reshaped after attending an online professional development course on MBT. A comparison of two different countries is particularly interesting since both countries have recently included modeling as a key component of their new national science curricula (Chile) and provincial science curricula (for example, British Columbia, Canada).

This research on assessment literacy among science teachers is potentially significant for the following reasons. Firstly, for researchers in assessment, the study presents a new instrument to identify ISTs' ALMBT in modeling. MBT is important in the science classroom not only because it helps students organize, reason, and visualize their understanding about systems or phenomena but also because the creation and use of models are core scientific practices that promote the development of reasoning skills in science. Therefore, assessment literate ISTs must be able to use their epistemological knowledge about a model to design and implement assessment practices to assess students' reasoning with models. Secondly, for the Ministry of Education in Chile, the study can generate data about how the new Chilean science curriculum is being implemented (including assessment) in schools in Chile. Similarly, the results of the administration of the questionnaire in Canada can be useful for Ministries of Education across

Canada to inform how ISTs assess students' models in the science classroom. Thirdly, the findings related to the second phase of this study open the way to identify teachers' strengths and weaknesses in their preparation and suggest what aspects of the dimensions related to their assessment literacy in MBT need to be enriched. It is worth noting that prior studies, such as Khan (2011a), have studied science teachers' strategies when teachers implement MBT. In Khan's study, the author observed that ISTs were not able to appropriately engage students in the evaluation and modification of models. This study was valuable as one of the first that identified what is missing in MBT. A key difference between this previous research and my study is that prior research has focused on identifying specific teaching strategies that ISTs use in their pedagogy to promote modeling whereas in my research, I build on the category of assessment practices as a pivotal component of teachers' pedagogy that could be further explored in MBT. An analysis of the influence of a 10-hour free online professional development course (OPDC) on ISTs' pedagogy and ALMBT was also explored. Professional development is important to help ISTs enrich and refresh their knowledge about how to teach science. For example, van Driel and Verloop (1999) stress the relevance of engaging science teachers in interventions that provide specific information and relevant literature in MBT to help them understand the predictive nature of models. It is well-established in the literature that many science teachers have "poor" and "confused" understanding about model and modeling (Danusso et al., 2010, p. 871) but also that their pedagogy does not always engage students in modeling practices or in the implementation of its various phases (Khan, 2011a). Prior studies have focused on analyzing the impact of MBT on the development of these modeling practices among students. Nevertheless, research in MBT could pay further attention to the role that teachers' assessment literacy plays in

guiding and advancing the generation, evaluation, and modification of models, as modeling is an important facet of science.

The objective of developing an OPDC in MBT had two main goals i) to offer support to ISTs about how to assess students in MBT, and ii) to gain deeper knowledge about the role of teachers' assessment literacy in supporting students in the generation, evaluation, and modification of their models, also known as the GEM cycle. This OPDC environment was different from experiences of professional development suggested by other authors (for example, Kawasaki & Sandoval, 2020) in the sense that the OPDC instructed ISTs in the foundations in MBT and suggested examples of assessment in models and modeling based on the research literature rather than focusing on helping teachers to collaboratively create plan lessons to include models in their pedagogy. Moreover, the duration of the online course was brief, approximately 10 hours, in comparison to other professional development interventions that usually cover at least a couple of days or weeks (c.f. Guy-Gaytán et al., 2019; Ogan-Bekiroglu, 2007). In my study, the OPDC was implemented within a month and each module was designed to guide teachers not only in the foundations of MBT but also regarding how to select models from the science curriculum and include them in their lessons, how to teach with models, and how to assess students' reasoning with models.

1.2 Overview of the Study

In this study, I contribute to our understanding of science teacher assessment, specifically as it relates to teacher's assessment of students' generation, evaluation, and modification of models. I explored science teachers' assessment literacy when students are engaged in reasoning with

models, an important aspect of science curricula world-wide. The analysis of data obtained from the administration of a close-ended questionnaire in MBT, and the detailed analyses of transcripts from interviews and class observations, contributed to responses to research questions 1 and 2. These questions asked if ISTs' knowledge of models and modeling was related to ISTs' assessment literacy in MBT and investigated in what ways ISTs' assessment literacy about models and models influenced their pedagogy.

Finally, this study also offers a portrait of the challenges that many science teachers have when teaching science using models. By exploring teachers' ALMBT from teachers across Canada and Chile and using in-depth observations of Chilean teachers, this research identifies and describes the dimensions related to assessments that teachers enact when teaching with models. Moreover, the study offers guidelines for curriculum developers and science teachers about clear criteria or aspects that they can focus on to enrich science teachers' assessment pedagogy.

1.3 Significance of the Study

Several studies have shown the limited knowledge that ISTs possess about models and modeling in science education. These studies have mostly focused on the difficulties that teachers possess in enacting MBT in their classrooms (see, for example, Justi & van Driel, 2005; Khan, 2011a); an even fewer number of studies have investigated how ISTs assess students in MBT. In this study, I characterize ISTs' ALMBT and explore how assessment literacy about models and modeling influence their pedagogy after participants attended the OPDC in MBT. Teachers' knowledge of assessment is key in the teaching-learning process in science education to ensure the successful implementation of new science curricula (Tacoshi & Fernandez, 2014). In this

vein, this study on teachers' knowledge of assessment is significant; it will contribute to the development of a deeper understanding of ISTs' ALMBT; and the findings might allow researchers and curriculum developers to identify potential difficulties that science teachers face in the implementation of a key practice in science education-modeling.

1.4 Organization of the Dissertation

This dissertation is organized into five chapters. The dissertation begins with i) the introduction to the problem in which I briefly review the importance of conducting research on assessment literacy in MBT and ii) an overview of the study. Chapter Two presents the theoretical framework and literature review. I elaborate upon the theoretical framework to define assessment literacy and I suggest some new elements to be included in the most current definitions of assessment literacy. Specifically, I emphasize scaffolding and learning progression as another element to be included in the definition of assessment literacy. Also, an introduction to the problem is presented where I argue the necessity to explore teachers' ALMBT. I include in this chapter the methods used to conduct the literature review on ALMBT.

Chapter Three presents the research design and methods for the current study. In this chapter, I provide details about the demographic information of the participants, and I describe the setting. Then, I describe the instruments used to collect data. I present the instruments for each phase of research, and the data collection and data analysis procedures. Finally, I explain the coding protocols used to analyze the qualitative data and discuss the issues pertinent to mixed methods research. Chapter Four presents the findings and discussion of these findings to answer the research questions. The findings have been organized into two sections, starting with establishing

a baseline of ISTs assessment literacy in MBT, followed by qualitative data on the identification and development of assessment literacy proficiency in MBT. Chapter Five draws conclusions for this study and makes assertions based on the cross-case analyses for ISTs' assessment literacy in MBT based. This final chapter concludes with implications for future research.

Chapter 2: Introduction to the Study on IST Science Assessment Literacy

2.1 Introduction

Teachers dedicate a significant part of their pedagogy to assessing students (Levy-Vered & Nasser Abu Alhija, 2015). That is, on average, nearly a quarter of teachers' pedagogy is spent on activities related to assessment (e.g., revising and modifying pre-developed assessments, providing feedback, verifying the quality of their assessment instruments) (Stiggins & Conklin, 1992). Hence, assessment constitutes an important knowledge base for teachers (Siegel & Wissehr, 2011), and enactment of this knowledge base can influence student outcomes. For example, the development and use of effective assessment techniques and grading practices can enrich students' levels of achievement when they are correctly implemented (Mellati & Khademi, 2018). Teachers' ideas about assessment help them make assessment decisions and adjust their instruction in the short- and long-term based on their students' results in the classroom (Abell & Siegel, 2011; Parr & Timperley, 2008; Siegel & Wissehr, 2011). In this sense, teachers should be "assessment literate" to effectively implement varied assessment strategies in the classroom to be able to make informed decisions about how to improve students' learning (Deluca et al., 2013). I conducted this literature review as a contribution to the limited research in assessment literacy in MBT as well as a review of the state of the field for context of study reported in later chapters. Taking into consideration the limited research in this area, the systematic approach used for the present literature review on assessment literacy, attempts to; i) identify the dimensions of assessment literacy included in studies in research on MBT; ii) enlarge the pool of known strategies by characterizing the methods researchers in science education used to investigate science teachers' assessment strategies in MBT; iii) describe and compare researchers' strategies implemented to enrich teachers' ALMBT; and iv) explore what

the findings from the systematic approach to the review informs us about science teachers' ALMBT, where possible, including areas to inform science teacher education.

2.2 Assessment Literacy

For the purpose of addressing science teachers' ALMBT, and as an advanced organizer for this section, I conceptualize assessment literacy based on research over the last three decades, and I examine the different dimensions or sub-constructs that have previously constituted assessment literacy. I also suggest a reconceptualization of assessment literacy in which I include a new dimension related to teachers' knowledge of learning progression and scaffolding. In this reconceptualization, I point out that when science teachers use models in their pedagogy, they must have a good knowledge of the nature and purpose of models and must also know how to implement assessment practices to guide students during the process of construction, evaluation, and modification of students' models. I also suggest that ISTs' assessment literacy may be different when a science teacher assesses core ideas in science that do not necessarily involve the use of models. For example, ISTs' assessment literacy to assess course-specific elements, such as chemical nomenclature, may differ from their assessment literacy when their pedagogy involves teaching with models and modeling.

2.2.1 Conceptualizing Assessment Literacy

The expression "assessment literacy" was not coined until 1991 by Stiggins (1991b). This scholar points out that assessment literacy is related to a teacher's capacity to understand whether assessment informs students about the achievement outcomes that a teacher values. Stiggins also states that assessment literacy relates to a teacher's capacity to identify, "When an assessment

target is unclear, when an assessment method misses the target, when a sample of performance is inadequate, when extraneous factors are creeping into the data, and when the results are simply not meaningful to them" (p. 535). From a pedagogical perspective, teachers' knowledge of assessment has been associated with the model of Pedagogical Content Knowledge (PCK) suggested by Shulman (1986). Even though Shulman initially emphasized PCK only as an amalgam of content knowledge and pedagogical knowledge, Magnusson et al. (1999) state that teacher knowledge of assessment is a key component of PCK and is connected to other types of knowledge that teachers must have, such as: curriculum, instruction, and students. This conceptualized knowledge of assessment includes two main dimensions related to i) knowledge of what to assess and ii) knowledge of how to assess in the classroom. When conceptualizing assessment literacy, we need to understand that our views of learning and teaching may shape this type of knowledge, for example, *traditional* versus *constructivist* views. In this sense, the role of assessment might change synchronously with such variations in the views of learning and teaching (Abell & Siegel, 2011). Delandshere and Jones (1999) point out that in a traditional approach, teachers use, "assessment of learning" to sanction and verify students' learning regarding specific content. In this case, assessment is traditionally implemented after learning to check what students have learned in terms of measuring the accumulation of knowledge (Abell & Siegel, 2011; Gottheiner & Siegel, 2012). From a constructivist perspective, teachers use assessment to provide evidence of how student thinking evolves (Ogan-Bekiroglu & Suzuk, 2014), help students compare and contrast their knowledge, and critique their own understanding (Koh, 2011).

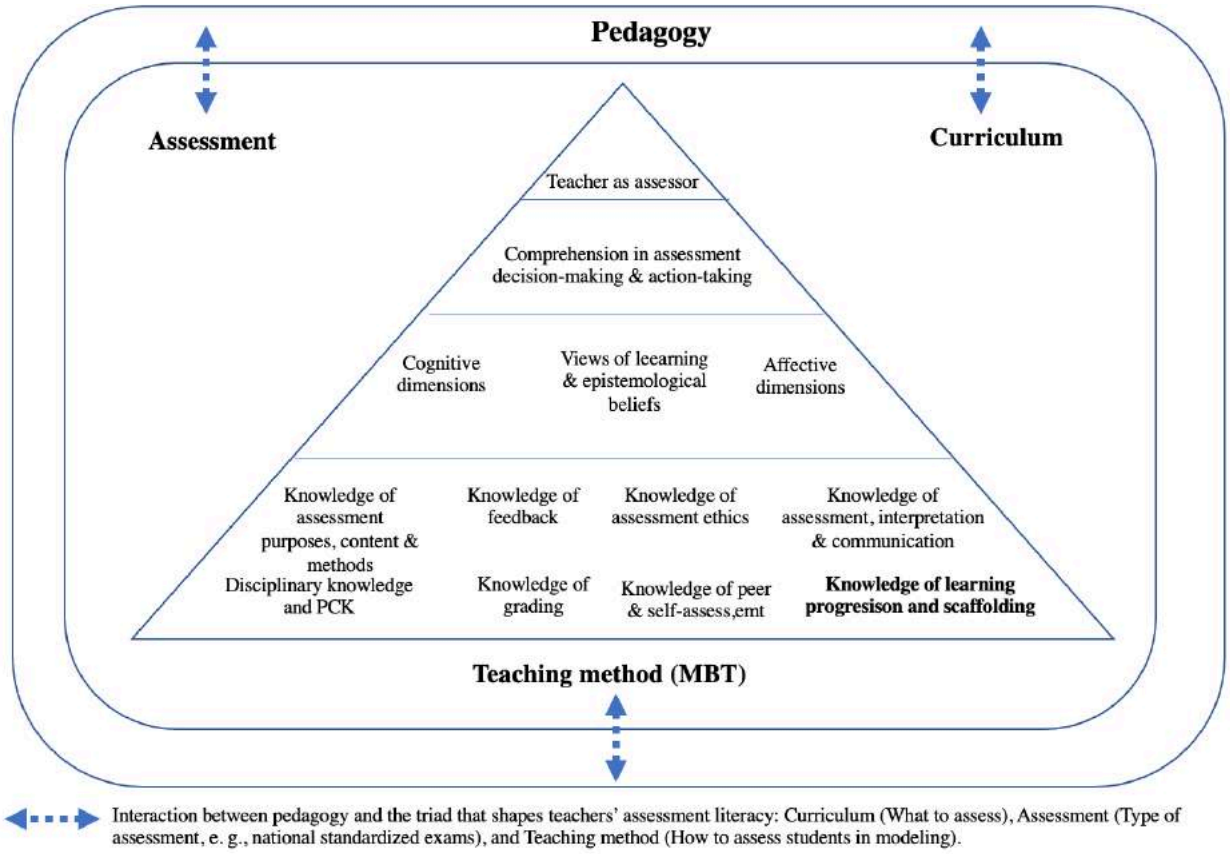
Another theoretical foundation to understand assessment literacy comes from the American Federation of Teachers, the National Council on Measurement in Education and the National Education Association (AFT, NCME, & NEA, 1990). These organizations developed the *Standards for Teacher Competence in Educational Assessment of Students* (*1990 Standards)*. These standards included seven areas to guide teacher preparation and establish what is understood to be an assessment-literate teacher. Examples of standards include; selecting appropriate methods of assessment, developing assessment methods that are appropriate for assessing students, and using assessment to make informed decisions. These standards have been updated by some scholars (e. g., Brookhart, 2011) and still remain as an important authority to guide teachers in their daily practice of evaluating students (Muhammad et al., 2020) and have been used to guide the elaboration of instruments to investigate levels of assessment literacy such as *Teacher Assessment Literacy Questionnaire* (TALQ) (Plake et al., 1993) and *Classroom Assessment Literacy Inventory* (CALI) (Mertler, 2004).

A more recent definition of assessment literacy comes from Xu and Brown's (2016) work which is based on the *1990 Standards* and introduces a new framework of teacher assessment literacy "in practice" (TALiP). Their conceptualization of assessment literacy not only includes the seven standards that constitutes the base of a pyramid of teachers' knowledge of assessment (see Figure 1), but also other "practice" domains. In their conceptualization, assessment literacy is shaped by a group of domains related to teacher assessment that include "*teacher conceptions of assessment"* (cognitive dimensions, views of learning, epistemological beliefs, and affective dimensions), "*institutional and socio-cultural contexts*" (different goals and outcomes based on different contexts), "*teacher assessment literacy in practice"* (compromises made in assessment,

decision-making and action-taking), "*teacher learning*" (e.g., enrichment of their understanding of assessment), and "*teacher as assessor*". This last domain refers to teachers' capacity to empower themselves with their own autonomy and various repertoires to assess students. Figure 1 shows my adaption of Xu and Brown's theoretical framework of assessment literacy.

Figure 1

Adaptation of Xu and Brown's Theoretical Framework of Assessment Literacy (2016)



Note: Only the new aspects included in the framework are bolded.

In my study, Assessment literacy in MBT (ALMBT) is defined as a *multidimensional construct that is comprised of a set of knowledge and skills about the assessment of models and modeling which is activated in the science classroom when reflecting on practice while students generate,*

evaluate and modify their initial models. I assume that ALMBT is built and enriched over time since teachers need to develop, test, and refine their assessment practices and comprehend their role as assessors when making pedagogical decisions about their assessment practices used to assess students' reasoning with models.

The conceptual framework that I take forward in this study is very similar to that developed by Xu and Brown (2016). I used Xu and Brown's conceptual framework of teacher assessment literacy in practice which is represented by a pyramid. In Figure 1, I show in bold type the adaptation of Xu and Brown's model of assessment literacy used in my study. Firstly, I state that assessment literacy is framed by teachers' ideas about pedagogy. For example, teachers who adopt a constructivist pedagogy must create situations in the science classroom in which students question their own and peer ideas and create knowledge through the interaction with others. In this sense, teacher' pedagogy might influence i) how teachers teach the prescribed curriculum (e.g., curriculum goals, selection and organization of content, use of learning progressions and scaffolding), ii) how teachers assess their students (e. g., traditional exams versus the assessment of generated models), and iii) the teaching method used to implement the curriculum and assess students (e. g., content-based lecture versus inquiry-based teaching). Secondly, I point out that assessment literacy is shaped by this triad (assessment, curriculum and teaching method). This new aspect of Xu and Brown's framework is depicted by the intersection among the three circles (assessment, curriculum, and teaching methods) included in Figure 1. When teachers develop and implement assessment strategies, there is an interaction and tension between curriculum, assessment, and teaching approaches (Abell & Siegel, 2011; Carr et al., 2000; Shepard, 2000). On the one hand, curriculum reforms affect teachers' assessment strategies and influence and

determine what to assess. For instance, in the Chilean science curriculum, it is expected that students, "Evaluate the validity and limitations of a model or analogy in relation to the phenomenon modeled" (MINEDUC, 2019, p. 50). In this curriculum, modeling is included as one of four skills and steps involved in the scientific process. Specifically, the Chilean science curriculum state that teachers should encourage their students to i) plan and conduct an investigation; ii) analyze and interpret data; iii) construct explanations and design solutions (argumentation, designing projects, and models), and iv) evaluate. Hence, assessment literate teachers in MBT must be able to promote and assess the construction of explanations and the design of solutions that involve the generation of models by incorporating in their pedagogy and assessment instruments the guidelines suggested by the national curriculum. On the other hand, specific types of assessment, such as standardized national testing might influence how teachers implement the prescribed curriculum to evaluate students' achievements. I assume that teachers can show different levels of proficiency in assessment based on the teaching methods or approaches they use to teach and assess the prescribed science curriculum (e. g., traditional approach; project-based teaching; model-based teaching; argumentation-based teaching). In other words, a teacher might be highly "assessment literate" when implementing a traditional method of teaching that focuses on measuring students' factual knowledge; however, s/he might have a limited ALMBT because of his/her lack of experience or knowledge about how to include models in his/her pedagogy and lack of repertoire to assess students when thinking with models. Hence, in MBT, I assume that this triad, which is not explicitly included in Xu and Brown's work, shapes what to assess from the curriculum (e. g., generation of a model of photosynthesis versus a predefined model), why to assess (e. g., modeling practices instead of rote learning) and how to assess (e. g., implementation of rubric performance tasks aligned to modeling practices).

Thirdly, I point out that teachers' knowledge of assessment needs to explicitly include teachers' competence to assess students thinking not only in the short-term but also in the long-term to facilitate students reshaping, enriching, and refining their ideas (*learning progression and scaffolding)*. I broaden the definition of assessment literacy suggested by Xu and Brown by including this new temporal dimension of learning because I believe teachers need to be able to assess how students' learning and modeling practices progress in the classroom over the long term, for example, by implementing assessment strategies that facilitate the identification of students' prior knowledge and promoting the modification or enrichment of students' initial ideas or models in subsequent periods.

In the following section, I describe each of the dimensions included in my conceptualization of assessment literacy to clarify each of the components. It is worth mentioning that in my study I mainly focus on the "basic mastery of educational assessment knowledge" (Xu & Brown, 2016, p. 158) related to each of the dimensions included in the base of the pyramid. The higher levels suggested in Xu and Brown's framework are tangentially explored even though they do not represent the main focus of this dissertation. These higher levels are related to teachers' perceptions of how assessment should be and their conceptions of assessment and their self-directed awareness of the assessment processes.

2.2.1.1 Disciplinary Knowledge and Pedagogical Content Knowledge

As was already detailed in Figure 1, the knowledge base for effective assessment is comprised of a set of knowledge that teachers need to master. Disciplinary knowledge and pedagogical content knowledge are one of the basic types of knowledge that teachers have to possess in order to

know what and how to assess students. My view of teacher knowledge for this dimension draws on areas of agreement in Windschitl's (2004) ideas that have originated from Shulman's (1986) original conceptualization of teachers' knowledge. In his classification of teacher knowledge, Windschitl distinguishes between four types of knowledge. The first aspect includes *general pedagogical knowledge* which is related to those strategies that teachers implement in their pedagogy within a classroom setting (e.g., being able to moderate discussions, knowledge about classroom management) (Borko & Putnam, 1996). The second aspect relates to *content knowledge* (e.g., understanding of disciplinary core ideas). The third aspect, called *pedagogical content knowledge*, considers teacher knowledge of how students comprehend the discipline, for example, understanding of alternative ideas. Finally, he suggests *disciplinary knowledge* as, "How knowledge was produced and judged in a particular domain of inquiry" (p. 5). Even though traditionally the understanding of the discipline has been included as a component of content knowledge (disciplinary core ideas such as domain's concepts, theories, and laws), Windschitl stresses out that in MBT, disciplinary knowledge corresponds to its own category (e. g., "understanding of the purposes of science inquiry", "understanding the nature of relationships between scientific models and data" (Windschitl, 2004, p.5). In my study, I also take this dimension for the knowledge base of assessment as the specific body of knowledge that is comprised of two types of knowledge. On the one hand, disciplinary knowledge in my study is understood as the knowledge about scientific models (e. g., nature and purpose). On the other hand, I point out that pedagogical content knowledge is understood as the knowledge about the role of models in science education and how students understand science while reasoning with models (how to engage students in model-making). For example, teachers with an unsophisticated knowledge about the nature of models might tend to implement assessment

practices that emphasize the memorization of isolated components, features or characteristics of a curricular model instead of promoting students' understanding of the purpose of model-based inquiry in the science classroom. In other words, ISTs not only need to possess an accurate epistemological knowledge of models (disciplinary knowledge) but also know how to promote modeling to help students understand the role of models to represent, communicate, analyze evidence, and explain and predict phenomena (PCK). For example, Pluta, Chinn, and Duncan (2011) advocate for the importance of assessing students' own epistemic criteria for good scientific models. In this context, beside their disciplinary knowledge about the nature of models, teachers must be able to involve students in thinking with and about models and also have enough knowledge and assessment repertoires to elicit and evaluate students' generated models and their modeling practices (Cheng & Lin, 2015a). Research has shown that many science teachers have a limited knowledge of models and modeling in science education (Crawford & Cullin, 2005) which further calls into question science teachers' knowledge of how to assess students when using models.

2.2.1.2 Knowledge of Assessment Purposes, Content and Methods

The goal of learning about assessment is not just to develop different tests to measure students' understanding of content. The adapted framework from Xu and Brown (2016) includes as another component of the knowledge base the dimension related to knowledge of assessment purposes, content, and methods. When teachers implement assessments, their purpose also covers i) the use of formative and summative assessment to guide students' learning), ii) provision for feedback, iii) gauging students' alternative ideas and prior knowledge, and iv) helping students develop new skills, among others. In this sense, assessment methods need to be

strategically designed to ensure that they are aligned with the curriculum and are adequate for particular students (Siegel & Wissehr, 2011). For example, summative assessments can be planned to measure student's learning at specific junctures once instructional events have been completed (Popham, 2009), whereas formative assessment can take place during instruction and be used to provide feedback to students as well as evidence of their engagement in the learning process (Abell & Siegel, 2011; Bennett, 2011; Deluca et al., 2013). When teachers engage in summative and formative assessment, they are able to adjust their pedagogy based on the evidence collected in the classrooms (Mellati & Khademi, 2018). Moreover, teachers need to be aware of the advantages and disadvantages of the assessment instruments that they choose to assess students (Siegel & Wissehr, 2011) when implementing a specific approach to teaching science.

2.2.1.3 Knowledge of Grading

Another theoretical dimension related to assessment literacy included in Figure 1 is related to teachers' knowledge of grading. This knowledge refers to teacher capacity to develop scoring techniques to assess their students, for example, for grading students' understanding and performance in class (e.g., criteria and rubrics). Popham (2009) emphasizes that teachers need to be able to construct, test and improve their test items. Nevertheless, Schafer (1993) states that assessment in the classroom is often implemented haphazardly and teachers might lack knowledge about how to plan, implement, interpret, and judge assessment when evaluating students' achievement. In the case of MBT, if teachers in the science classroom ask their students to carry out an investigation, or create and refine a model, it would be reasonable to expect that summative assessments would also focus on assessing modeling and inquiry skills

instead of assessing rote learning (e.g., asking to define concepts or find the correct definition in a multiple-choice question). Therefore, teachers need to develop assessment instruments in which they clearly identify and measure students' learning goals (Stiggins, 1995) by aligning assessment with instruction.

2.2.1.4 Knowledge of Feedback

Xu and Brown also emphasize that teachers must be able to support, enhance or assist student learning during assessment (Gan et al., 2018; Wakefield et al., 2014). As indicated in Figure 1, teacher knowledge of feedback is one of the components of the knowledge base in assessment. By providing feedback, teachers help students understand what is expected (Stiggins, 1991a) and this may facilitate students' self-regulation (Carless, 2006), or awareness of their own learning process. Hattie and Timperley (2007) point out that feedback in the classroom can include, for example, feedback about: the task, the processing of the task, self-regulation, and the self as a person (emotions and affections). In the process of teaching and learning, the quality of the feedback is also important. Feedback can, "[I]nvolve a comparison between [:] the student's achievement or performance and other students' (norm-referenced), standards or learning goals (criterion-referenced), or the student's previous achievements" (Bell, 2007; p. 977). The moment in which feedback occurs is another important factor to consider. For example, feedback at the end of the assessment process does not always offer students enough learning opportunities to improve their performance (Yorke, 2001). Hence, one of the main goals of feedback is to contribute to closing the gap that students experience regarding the ideal performance that they should achieve in class and their current performance (Zhang & Zheng, 2018).

2.2.1.5 Knowledge of Assessment Interpretation and Communication

Interpretation and acting upon assessment data are also crucial components of assessment literacy (Abell & Siegel, 2011) since teachers need to use the evidence from assessment to help them judge student achievement and enrich their pedagogy (Egan & Archer, 1985). Figure 1 depicts this component of knowledge of assessment interpretation and communication in the base of the triangle. Popham (2009) states that some teachers tend to focus assessment methods to solely grade students' performance rather than using the evidence from assessment to verify students' understanding or to make inferences about what students have or have not learned. The interpretation of the findings from the assessment allows teachers to shape and inform their own pedagogy as well improve students' learning and report progress. Moreover, it can be asserted that teachers' capacity to accurately assess student achievement and interpret evidence cogently depends on teachers' pedagogical content knowledge about how to teach the subject (Shulman, 1986) and their own experiences and training, professional development, and personal backgrounds (Edwards, 2016; Martínez et al., 2009). In this sense, an assessment literate teacher in MBT must be able not only to judge students' understanding of the phenomenon to be modeled but also identify how the modeling process occur in order to adjust his/her instruction based on the results obtained from the assessment.

2.2.1.6 Knowledge of Peer and Self-assessment

An assessment literate teacher also needs to be able to use assessment to motivate student learning, according to authors Wanner & Palmer (2018). Xu and Brown (2016) state that teachers need to master assessment practices to train students to effectively participate in an assessment. In this vein, Boud (2000) further states that assessment, "[H]as to move from the

exclusive domain of assessors into the hands of learners" (p. 2000) which can be interpreted that students need to have an active role and be engaged in assessment. Fredericks et al. (2004) and Munns and Woodward (2006) point out that by engaging students in assessment thought self-assessment, teachers can benefit and i) reflect on what they (students) have learned, ii) evaluate students' performance, iii) see the classroom as a community of learners, iv) value themselves [students] as individuals and learners, and v) be part of the learning experiences and the assessment process. While there are many ways to engage students in assessment, two of the considered approaches are: *peer* and s*elf-assessment*. On the one hand, self-assessment is related to students' capacity to take responsibility of their learning process in which they reflect, adjust, and judge their own ideas or performance (Willey & Gardner, 2010). On the other hand, peer assessment corresponds to those activities that students conduct to make judgments about others' work (Reinholz, 2016). This process of peer assessment has potential to contribute to constructive collaboration in the process of learning (Hanrahan & Isaacs, 2001). In either case, I hypothesize that an assessment literate teacher should be able to use self- and peer-assessment to support teaching and peer learning from a formative and summative perspective to help students understand how the process of assessment occurs in the classroom.

2.2.1.7 Knowledge of Assessment Ethics

Another component of the knowledge base of assessment literacy suggested in Figure 1 is related to teacher knowledge of assessment ethics. Gipps (1994) emphasizes that, "[A]ssessment is a powerful tool: it can shape curriculum, teaching, and learning; it can affect how pupils come to see themselves both as learners and in more general sense as competent or not; through labelling and sorting pupils (certificating and selecting) it affects how pupils are viewed by others" (p.

144). Xu and Brown (2016) emphasize that assessment literate teachers need to understand the ethical principles of assessment in which the use, storage, and distribution of assessment results need to be considered to improve students' learning and promote inclusion and equity of access in class. Since MBT involves a process of co-construction of models through the interaction between the teacher and the student, I further suggest that teachers' assessment strategies need to offer each student the equitable opportunities to express, revise and evaluate their models and explanations.

2.2.1.8 Scaffolding and Learning Progression

As previously stated in Figure 1, In my conceptualization of ALMBT I included scaffolding and learning progression as an important component of this knowledge base. The reason for including this component in my adaptation of Xu and Brown's framework of Assessment Literacy (Figure 1) is because a challenge in science education has emerged over the last years regarding how to assess students learning progression in the science classroom, and particularly in MBT (see, for example, Schwarz et al., 2009). More traditional approaches consider learning progressions as a guideline developed by content experts or curriculum developers to describe how content accumulates and gets more complex when students' progress through a course or unit (Alonzo, 2018). In science education, and in my study, learning progression is understood as a tool for assessment and learning that is based on research and teachers' experiences about how student thinking evolves (Shepard, 2018). Moreover, learning progressions are not comprised of fixed linear pathways (Pierson et al., 2017) and require an iterative refinement process to establish. Teachers should not merely think of scaffolding as an instructional approach to help students advance their understanding of different phenomena or disciplinary core ideas in

science. Instead, science teachers must value the role of assessment as a tool that involves a collaborative process of gathering and analyzing evidence through assessment by negotiating the meaning during the student and teacher interaction in the science classroom while students develop, test, and use a model. In this collaborative process, by analyzing students' reasoning with a model, assessment literate teachers need to be able to design formative and summative assessments to identify students' modeling performance expectations at multiple points along their learning progression and align the development of these modeling practices with the understanding of the expected curricular model.

In the science classroom, learning progressions are especially used as support for teachers' formative assessment practices because these progressions allow teachers to identify "the gap between students' current understanding and what is targeted, thus, what could be done to support students in closing that gap" (Alonzo, 2018, p. 109). Hence, the Council of Chief State School Officers (CCSSO) in the United States acknowledges learning progressions as one of the five attributes of effective formative assessments (McManus, 2008). The Next Generation Science Standards also includes them as a key dimension in the science content standards (NGSS Lead States, 2013) and as a key component of the eight Science and Engineering Practices (Pierson et al., 2019). For example, Appendix F in the NGSS for grades 6-8 for the practice related to modeling details that students should be able to, "Evaluate limitations of a model for a proposed object or tool", whereas in grades 9-12, it is expected that students "Evaluate merits and limitations of two different models of the same proposed tool, process, mechanism or system in order to select or revise a model that best fits the evidence or design criteria." (p. 2). This brief example of a learning progression suggests that assessment literate teachers must be able to

guide and assess students learning based on how their ideas of how model development progresses in complexity. Schwarz et al. (2012) emphasize that teachers should be able to inform teachers and help them differentiate between learning progressions that facilitate *the development, improvement or enrichment of learning*, and those teacher assessment practices that inform learning progressions to enhance students' *scientific practices*. Gotwals (2018) points out that in both types of learning progressions, teachers must challenge students to apply their understanding across different contexts to make appropriate judgments and interpretations of students' level of understanding.

To summarize, I defined assessment literacy as the knowledge and skills about the assessment of models and modeling and the decision-making process that teachers activate when teaching and assessing models and modeling practices in the science classroom. This knowledge base is shaped by the theoretical dimensions suggested by Xu and Brown (2016). In addition to the dimensions suggested by Xu and Brown (2016) (i) disciplinary knowledge and PCK, ii) knowledge of assessment purposes, content, and methods, iii) knowledge of grading, iv) knowledge of feedback, v) knowledge of assessment interpretation and communication, vi) knowledge of peer and self-assessment, and vii) knowledge of assessment ethics, I included a dimension related to learning progression as a key component of ALMBT. Therefore, ALMBT can be investigated in teachers' pedagogy based on the identification of each of these dimensions in practice. Each of the theoretical dimensions comprise assessment literacy and will be used to orient the forthcoming review of literature on ISTs' ALMBT for this study. The following section presents the problem that guides the review of literature and the methodology used to conduct it.

2.3 Literature Review Methods

2.3.1 Literature Review Approach

I utilized a systematic approach for a review of literature on assessment literacy in MBT, to be conducted spanning the last 30 years. These years were selected because modeling was announced as a promising area to research in science education in K-12 classrooms roughly three decades ago (Clement, 1989; 1993; 2000) and more recently, there has been a proliferation of studies on MBT (Beck et al., 2020; Buckley et al., 2004; Campbell et al., 2011; Clement, 2000; Danusso et al., 2010; Gilbert, 2016; Guy-Gaytán et al., 2019; Justi, 2009; Kawasaki & Sandova, 2020; Khan, 2007, 2011a, 2011b; Nunez-Oviedo & Clement, 2019; Ogan-Bekiroglu, 2007; Windschitl et al., 2008). The guiding questions that informed my review of the literature on assessment in MBT included:

1) What strategies have researchers and teachers used to assess students' models in science classrooms?
2) What strategies have been implemented to enrich teachers' understanding of MBT and assessment literacy in MBT?
3) What do the findings from a review of literature tell us overall about science teachers' assessment literacy in MBT?

To answer the guiding questions detailed above, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Moher et al., 2009) were used to select the articles and conduct the systematic review. PRISMA is a guideline for systematic reviews and meta-analyses of empirical studies widely used in the applied sciences and is applicable across

different disciplines. Even though there are other citation strategies, such as citation chaining or snowballing (Sayers, 2007), that involve using the reference list to trace an idea by checking forward and backward the sources cited in an article, I preferred a more rigorous approach to conduct the revision of literature. Moreover, using citation chaining would have narrowed the number of scholars and the diversity of lines of studies related to MBT and might have missed important literature. By using the PRISMA protocol, I was more flexible during the literature review to identify new authors that might emerge during the database search. Nevertheless, I acknowledge that snowballing sampling can be particularly useful for extending and complementing a systematic review. PRISMA is often used in science education. For example, Vojíř and Rusek (2019) used PRISMA to depict the process of paper selection for their review of science education textbook research trends (see also, Crompton et al., 2016; Margot &Kettler, 2019). Crompton and colleagues (2016) used PRISMA guidelines also to conduct a systematic review of the use of mobile learning in science. From an initial search of 1,532 articles, and after following the PRISMA protocol (e.g., screen abstracts, screen full text), only 49 studies were included in their final article selection.

The PRISMA guidelines are comprised of a checklist to assess published protocols and the appropriateness of methods (Moher et al., 2015) and its aims are "to reduce the risk of flawed reporting of systematic reviews and improve the clarity and transparency in how reviews are conducted" (Liberati et al., p. 22). I used this protocol to ensure the transparency of the results and conduct a rigorous analysis of the field of assessment of MBT. It is worth mentioning that it has been suggested that two or more reviewers be used to conduct the screening of articles against inclusion and exclusion criteria to avoid bias while using PRISMA (Crompton et al.

2016, Margot & Kettler, 2019). I acknowledge that using two or more reviewers might decrease the possibility of rejecting important articles when screening published studies; however, PRISMA does not mandate the use of two reviewers as an essential requirement and states double screening as a suggestion for researchers (Mahtani et al., 2020). Waffenschmidt and colleagues (2019) point out that a single screening can also be an appropriate methodological shortcut in rapid reviews and might still be useful to conduct a literature review. For this thesis, this methodological choice was carried out because a second screener was not available to conduct a conventional double screening; however, PRISMA was still considered to be a rigorous approach to identifying articles for my study.

In terms of inclusion and exclusion criteria, all of the articles for this review of literature had to meet the following inclusion criteria:

- Empirical Research: Empirical qualitative and quantitative research studies on MBT related to the phenomenon of interest (assessment practices that teachers or researchers have implemented to assess students' models).
- K-12 and University Teachers and Students: Studies had to involve science teachers or middle and secondary students whose main focus was on modeling practices in science education. Studies with university students and university teachers were also included if the purpose of the MBT intervention was to foster students' modeling practices and assess their constructed models or metamodeling knowledge (knowledge about the nature and use of scientific models and modeling in science).

- Detailed Methodology: To be included in the literature review, studies had to provide appropriate indications of the methodology used in the study (e.g., type of intervention, study context, data collection, analysis).
- Peer-review Articles: Only scholarly peer-reviewed journals were selected to ensure that the articles passed minimal scientific quality (Potvin & Hasni, 2014).
- Only Articles Published in English: This criterion was included since the vast majority of articles are published in English and are included in the largest database for peer-reviewed journals.
- Scope and Purpose: The articles were selected according to the relevance of the topic of interest. For instance, the articles had to answer the questions related to who (science teachers), what (assessment of students in science), and how (MBT) at minimum.
- Authors: Even though theoretical articles were not initially included in the literature review, works by certain prominent or key authors who have conducted studies in MBT were included to help explain assessment practices in terms of theory.
- Informative articles on assessment in MBT: I included studies from book chapters that explicitly implemented professional development courses, included activities to explore ISTs' modeling practices, or administered a tutoring intervention by researchers to elicit some students' mental models in after school contexts (for example, Rea-Ramirez, Clement & Nunez-Oviedo, 2008).

The exclusion criteria included:

- Publication Date Range and Relevance: I only considered research published within the last 30 years, after MBT was announced as a promising area of research in science education (Clement, 1989; 1993; 2000).
- Theoretical articles: Conceptual and theoretical articles that focused on teachers' and students' knowledge of models without emphasizing MBT or implementing a modeling intervention were excluded from this research (for example, Rinehart et al., 2016).
- Curriculum documents: These types of documents were not considered in the literature review since the main purpose was to identify peer-reviewed strategies that have been used to help students elicit their mental models and explore how teachers evaluate their students empirically.

2.3.1.1 Databases

The Education Resources Information Centre (ERIC: http://www.eric.ed.gov/), Google Scholar, SpringerLink, Taylor & Francis Online and Wiley Online Library were the main databases used for selecting articles. These databases are some of the most popular and complete indexing systems for research in education (Hasni et al., 2016). The following search terms were used in the databases: 'model-based teaching' OR 'model-based inquiry'; 'assessment AND scientific modeling'; 'teacher AND assessment AND scientific model'; 'learning progression AND scientific modeling'. I also included the words 'model-based inquiry' (MBI) as an equivalent to 'model-based teaching' since some authors use them interchangeably. An additional search was conducted within these databases and journals after replacing the word 'modeling' with the British spelling 'modelling'. A first search was conducted in July 2019. A second data search was conducted in March 2020 to include new articles published from July 2019 to March 2020.

Table 1 below shows the number of returns for each keyword included in the literature search, whereas Figure 2 shows the process of paper selection following the PRISMA method.

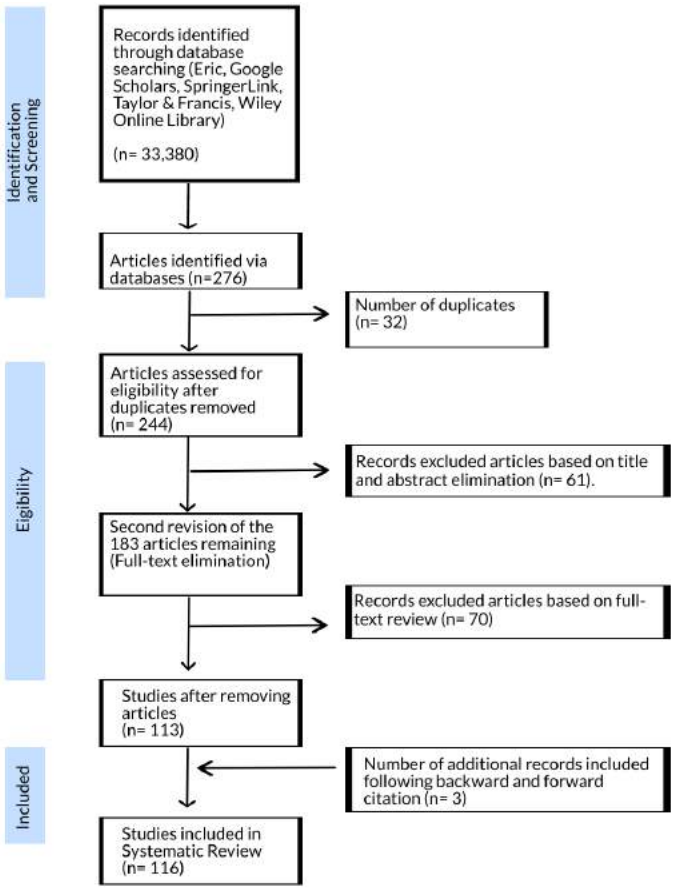
Table 1

Number of Returns for each Keyword Included in the Literature Search by Database

Keywords	Google Scholar	ERIC	SpringerLink	Taylor & Francis Online	Wiley Online Library
Model-based teaching	1407	172	154	39	15
Model-based inquiry	1976	360	136	38	47
Assessment AND scientific modeling	6341	154	3822	304	303
Learning progression AND scientific modeling	1061	4	143	42	34
Teacher AND assessment AND scientific model	14510	10	1201	606	501

Figure 2

A Diagrammatic Representation of the Literature Search and Review Process



The number of returns for each data search was merged in the table (both 2019 and 2020). In total, 33,380 returns were obtained from the data searches. The number of articles eventually included in the review of literature is shown in the following diagram (116) and detailed in the following section.

2.3.1.2 Articles for the Review

From the first search conducted in July 2019, 220 articles were retrieved based on the entire combination of search terms and databases referenced above. This procedure was repeated for the second literature search conducted in March 2020. After the second literature search, 276 articles were identified in total (220 and 56 articles, respectively for each search). Because the search was conducted in two different time intervals and the same keywords were used in different databases, there were duplicates from the retrieved articles which were removed from the initial number of articles (276). After conducting a single screening of the title of the total number of returns (title elimination), 32 duplicates were removed, and 244 articles were subjected to abstract elimination. In a first round, where the abstracts of the articles were read and the articles were skimmed, 61 articles were excluded because the studies were not aligned to the guiding questions of the literature review. Then, in a second review, the remaining articles (183) were assessed based on a full-text elimination. Seventy articles were excluded because i) the main focus was on an individuals' mental models about core ideas in science and did not provide enough examples of modeling practices; ii) the main focus was on individuals' (students or teachers) knowledge of the nature and purpose of models but did not emphasize modeling practices or how teachers implement MBT; iii) the topic of MBT was evident but there was not a thick description of how modeling practices occurred or how students were assessed in the

science classrooms; iv) the article provided assessment strategies related to modeling practices but the methodology of the study was not clear or the details of the assessment strategies implemented by teachers or used by the researcher were vague; v) the article suggested activities in MBT from a theoretical perspective but did not include an intervention, or vi) the main topic was not MBT in science. Hence, the number of studies that were included in the overall revision of studies, because they met all criteria, was 113. It is worth mentioning that three articles were added after checking reference lists to find additional studies for the review of literature (Bielik et al., 2020; Passmore & Svoboda, 2012; Spier-Dance et al., 2005) for a total of 116 articles) (see appendix A to review the full lists of articles selected and excluded in the literature review).

2.3.1.3 Analysis of the Literature and Coding of Articles

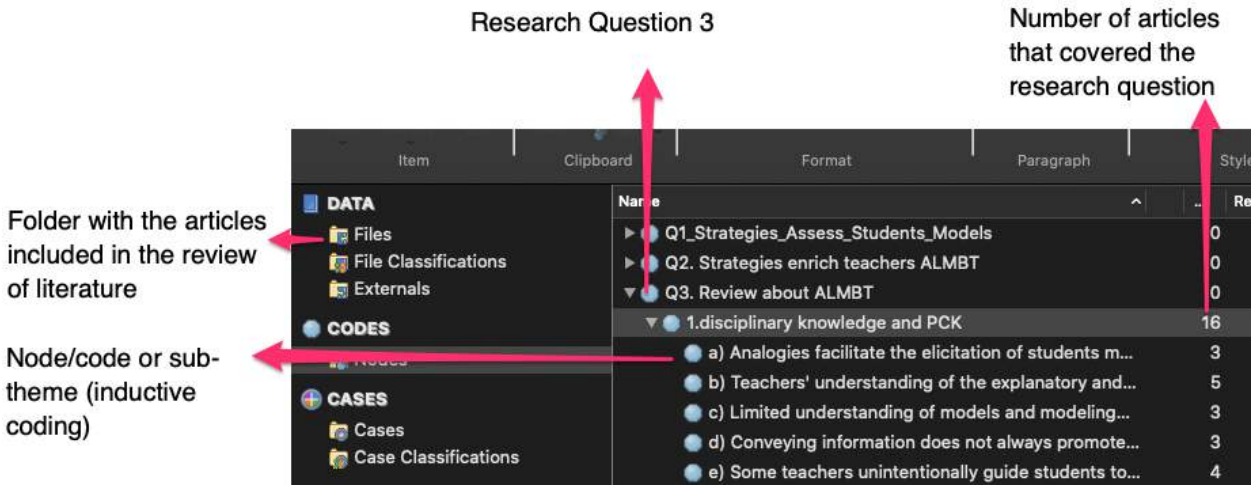
Once the articles for the review of literature were selected, a thematic analysis was undertaken. *Thematic analysis* (Braun & Clarke, 2006) was used as a method to investigate patterns within the articles. I will use "theme" consistently to describe specific patterns that captured the main ideas across the articles (Braun & Clarke, 2006; Xu & Zammit, 2020), as suggested by Sharp and Sanders (2019) who state that, "Thematic analysis is a qualitative method for identifying, analyzing, and reporting patterns within a data corpus" (p. 117). This analytic method has the advantage that it can be used to answer a wide variety of research related questions because it allows the researcher to identify, analyze, organize, describe, and report themes explored in different pieces of a data set (Braun & Clarke, 2006), and it is especially useful when analyzing large qualitative data sets (Nowell et al., 2017), including papers.

I started the analysis of the literature by applying inductive coding in which *open coding* was also implemented as a process that involved applying codes that emerged from the study, also called emergent codes (Blair, 2015). In this type of coding, the codes emerged from the data within studies (Xu & Zammit, 2020) and was used as a process of coding the studies to answer each literature review guiding questions without trying to fit the codes into a pre-existing coding frame (Nowell et al., 2017). In the case of the guiding question 3, related to what the findings from the overall review tells us about science teachers' assessment literacy in MBT, I also used induction¡ to locate gaps in the literature (Charmaz, 2006; Fereday & Muir-Cochrane, 2006) but the codes were organized based on the theoretical dimensions used to define assessment literacy (see Figure 3). Once the codes were identified, they were grouped into descriptive themes as is suggested by Ryan and colleagues (2018). Descriptive themes are broad and general codes that "capture and describe patterns in the data across studies" (p. 5). For instance, for guiding question 2, an example of a theme included "Professional development can help teachers to understand the foundations of MBT and enrich their assessment strategies". Examples of sub-themes for the overarching theme, based on open codes, included, for instance, "implementation of professional development to enrich teachers' assessment strategies" and "short-term professional development events". A thematic analysis then followed and was conducted using NVivo as a computer-assisted qualitative data analysis software (CAQDAS). This software program was used to organize the large number of articles. Zamawe (2015) points out that the presence of nodes in NVivo is especially compatible with thematic analysis approaches. A node corresponds to the assignment of a label to an excerpt of the data (in this case an excerpt from a published article). These nodes can also be merged through hierarchical coding in which the codes are organized and categorized based on the levels of specificity (King, 2004). Creating

node hierarchies makes it easier to identify the themes that emerge from the data. An example of the coding using NVivo to answer guiding question 3 for the literature review is shown in Figure 3.

Figure 3

Example of Coding for Guiding Question 3 for the Literature Review



The figure indicated above shows an example of coding for guiding question 3 for the literature review. In the top left corner is a reference to the folder with the files used in the literature review. In the middle section of the figure, the theoretical dimensions used to define ALMBT are included (e.g., 3. Disciplinary knowledge and PCK). These formed overarching themes (a theme that organizes and captures a larger number of sub-themes) and were only used to respond to guiding question 3 and summarize the findings. The nodes or codes that emerged from inductive coding correspond to sub-themes that were grouped to form the overarching theme. For example, for guiding question 3, for the theoretical dimension of disciplinary knowledge and PCK, some examples of sub-themes included “analogies facilitate the elicitation of students models”; “teachers’ understanding of the explanatory and predictive power of models”; “limited understanding of models and modeling and limited implementation of modeling practices”; and

"conveying information does not always promote modeling practices among students". Finally, the right column of the figure shows the frequency or the number of articles in which that specific dimension and sub-theme was uncovered. It is worth noting that the number of articles included for each of the themes does not necessarily match the number of studies included in the sub-themes. This difference in the number of articles is because most of the articles were coded more than once since, for example, researchers sometimes used multiple strategies to assess students' models (e.g., interviews and drawings). Therefore, on some occasions, the same article was included in more than one sub-theme, and it was used to answer more than one guiding question.

2.3.2 Review of Literature on Assessment Strategies in Science Education and MBT

This section answers the guiding questions formulated to assist in the review of the literature. The full list of the articles is detailed in Appendix A. As it was already mentioned earlier, the number of articles included by sub-theme do not match the number of total articles included for each guiding question because each article was coded more than once since they covered a variety of topics that were related to different sub-themes and guiding questions.

2.3.2.1 Guiding Question 1: What Strategies have Researchers and Teachers Used to Assess Students' Models in the Science Classroom?

Over the last thirty years, there has been a proliferation of literature in the field of model-based teaching (MBT). For example, I found eighty-six articles that were directly related to this

guiding question. From the existing research in MBT I reviewed, the main themes identified will be discussed below with the extant literature.

i) Class Discussion with the Teacher Helps Students Share, Compare and Evaluate their Models

The elicitation of models is a process that often requires interaction in the science classroom among students and the teacher. The articles included in this theme were related to the co-construction of a model in which students interact during the creation of a model in a process in which the teacher is a facilitator of model construction. Seventeen articles covered this theme (c. f., Aksit & Wiebe, 2019; Baumfalk et al., 2019; Brady et al, 2015; Jenkins & Howard, 2019; Justi et al., 2019; Kenyon et al., 2011; Khan, 2007, 2011a; Mendonça & Justi, 2011, 2013; Nunez-Oviedo & Clement, 2008; Oh, 2019; Peel et al., 2019; Rea-Ramirez et al., 2008; Ryu et al., 2015; Samarapungavan et al., 2017; Schwarz et al., 2009). In ten of the articles, consensus models were used as a strategy to assess students' models collectively (c. f., Baek & Schwarz, 2015; Baumfalk et al., 2019; Hokayem & Schwarz, 2014; Justi et al., 2009; Kenyon et al., 2011; Maia & Justi, 2009; Mendonça and Justi, 2011; Khan, 2007; Nunez-Oviedo & Clement, 2008; Ryu et al., 2015; Schwarz et al., 2009). Studies related to the co-construction of models are particularly well represented in studies conducted by Nunez-Oviedo and Clement (2008) and Khan (2007, 2011). For instance, Khan (2007; 2011) has explored teaching strategies that science teachers implement to involve students in chemistry classrooms. Her research has shown that the role of teachers during class discussions can promote cyclical patterns in which students can generate, evaluate and modify their models and/or hypotheses (Khan, 2007). Through this dynamic process of teacher- student interactions, the role of teacher is crucial for student

modeling. By promoting discussion with ideas that are contradictory to each other, teachers can assess students' initial models. In a study with middle-school students who attended small group tutorials after school, Nunez-Oviedo and Clement (2008) examined their discussions on models of the respiration process with the researcher. The teacher engaged students in discussions to contradict and evaluate others' ideas. During the competition mode, students were able to "display or express to the class two or more competing ideas at a time, providing an opportunity for comparisons (and therefore dissonance) before closure is reached on an idea" (p. 118). This competition strategy allowed the researcher to evaluate and detect students' ideas about a topic and helped students evaluate their own ideas and their peers' by providing opportunities for dissonance in their understanding of a topic. Ryu et al. (2015) engaged students in full-class discussion about surface tension and evaporation. In class, students' ideas were assessed based on how students identified conceptual issues, negotiated meaning, tested and modified their proposed ideas, applied concepts in their models of intermolecular forces, and reached agreement on concepts and statements. For example, the teacher guided students in the discussion of their models regarding how they could explain the relationship between the intermolecular force and the boiling points for two different substances such as water and ethanol. This group of studies further points out that the co-construction of models is a useful formative assessment strategy that teachers can use to engage students in the process of model evaluation.

ii) Assessment of Modeling Practices and Epistemic Aims of Models and Modeling

The second theme refers to the process of using assessment to engage students in modeling practices. Forty articles focused on the assessment of modeling practices, and they included four main subthemes; i) identification and modification of variables in students' generated models (22

articles) (Aksit & Wiebe, 2019; Bamberger & Davis, 2013; Campbell et al., 2012; Dauer et al., 2019; Demirhan & Sahin, 2019; Dickes et al., 2019; Fretz et al., 2012; Heijenes et al., 2018; Hernández et al., 2015; Khan, 2008a; King et al., 2019; Lally & Forbes, 2019; Peel et al., 2019; Pierson et al., 2020; Reinagel & Bray Speth, 2016; Samarapungavan et al., 2017; Schwarz et al., 2009; Spier-Dance et al., 2005; Sung & Oh, 2018; van Jookingen et al., 2019; Willensky & Reisman, 2006; Zangori et al., 2016); ii) identification of explanatory and predictive power to students' models (17) (c. f. Buckley et al., 2010; Dauer et al., 2019; Dickes et al., 2016; Duncan et al., 2016; Forbes et al., 2019; Fortus et al., 2016; Hernández et al., 2015; Khan, 2008a; Lally & Forbes, 2019; Peet et al., 2019; Pierson et al., 2020; Schwarz et al., 2009; Sung & Oh, 2018; Svoboda & Passmore, 2013; Vergara-Díaz et al., 2020; Wilkerson-Jerde et al., 2015; Zwickl et al., 2015); iii) model evaluation and evolution (21) (c.f. Aksit & Wiebe, 2019; Cisterna et al., 2019; Dolphin & Benoit, 2016; Forbes et al., 2019; Fretz et al., 2002; Heijenes et al., 2018; Hernández et al., 2015; Jong et al., 2015; Khan, 2007, 2008a, 2008b; Louca & Zacharias, 2015; Nunez-Oviedo & Clement, 2008, 2019; Reinagel & Bray Speth, 2016; Ruppert et al., 2019; Samarapungavan et al., 2017; Schwarz et al., 2009; Sung & Oh, 2018; Wilensky & Reisman, 2006; Zwickl et al., 2015); and iv) assessment of students' understanding about the nature and utility of scientific models through their meta-modeling knowledge of models and modeling (6) (Bielik et al., 2020; Chang & Chang, 2013; Fortus et al., 2016; Louca & Zacharia, 2015; Pluta et al., 2011; Prins et al., 2011). For instance, in Khan's (2007; 2008a) studies, the researcher investigated the instructional strategies that a teacher used to foster students' modeling practices and enrich students' mental models. Some of the strategies that the teacher used to assess students' understanding of models of intermolecular forces included fostering an inquiry process of making predictions and modifications of hypotheses. The teacher asked students to select

variables, provided discrepant information, and asked students to revisit their initial relationships within their models. Through this process, it was observed that the teacher involved students in dialogue of developing models as tentative constructions that can be revised to evaluate their logical, empirical, or conceptual consistency. In another example, Prins, Bulte, and Pilot (2011), in a study with secondary students in a chemistry course, observed that five strategies were implemented in class to assess and facilitate students' expression of their epistemological views of models and modeling. These teaching and assessment strategies included, for example, i) the visualization and conceptualization of anchoring problems in which students identified variables and recognized the purpose of the phases of modeling; ii) engaging students in the nature and the characteristics of their models to check the validity of the model and; iii) challenging students to use their constructed models in real-world settings. Sung and Oh (2018) also adapted three modeling practices to assess two sixth-grade science classes. In their study, the teacher included in the lesson plans three assessment practices during modeling that were related to epistemological reasoning. These epistemological practices included i) *expressive modeling* (students expressed their ideas by constructing and manipulating a model); ii) *experimental modeling* (students formulated hypotheses or predictions and tested their models by collecting data from experimental results); and iii) *evaluative modeling* (students compared alternative models to assess the scope and limitations of their models). By promoting these practices, students were assessed based on how they used their models as sense-making tools to explain their models of the seasons (e.g., "Being able to construct a physical model to express their ideas of the seasons", "Capability of producing the seasons with a physical model", and "Ability to revise a model in alignment with the scientific understanding of the seasons when being assisted by more capable others" (p. 857).

In a study with seven undergraduate students, Svoboda and Passmore (2013) explored the role of modeling strategies in biology education. These researchers assessed students' models by identifying five major pragmatic uses for models such as unrealistic models help scientists explore complex systems, models have predictive and explanatory power, and models can lead to the development of conceptual frameworks. While developing and working with models, students created and revised models in which they identified factors that could explain a vaccination-disease dynamic. By conducting thought-experiments and developing mathematical expressions, students used their models, speculated on results, found missing elements in their models, decided the main factors that could explain the phenomenon, and used their models to answer empirical data. The researchers assessed students' modeling practices by identifying students' capacity to i) accurately identify variables in a model and assess the explanatory and predictive power of their revised models, ii) and assessed their theoretical contributions of each model. Overall, the literature reveals that by challenging students to use their model and particularly by helping them understand the explanatory and predictive power of their generated models, researchers and teachers can assess students' models by promoting some major modeling practices such as asking students to generate, evaluate and modify their relationships between variables.

iii) Analogical Reasoning as a Tool to Mediate and Assess the Generation and Revision of Students' Models

The third theme refers to students' engagement in the generation of analogies to facilitate the understanding of curricular models. I found a total of nine articles that included analogies as one of the main strategies to help students think with models and assess their reasoning with a model

(Cuperman & Verner, 2019; Gobert, 2000; Han & Kim, 2018; Lee, 2015; Shemwell & Capps, 2019; Prins et al., 2011; Spier-Dance et al., 2005; Vergara et al., 2020; Yeşiloğlu, 2019). By generating analogies, students can compare models, which allows them to identify the limitations and strengths of competing models. The assessment of students' analogical reasoning was included in some articles as a strategy to evaluate how students thought with their models to represent and compare their ideas with a target phenomenon or model (see, for example, Yeşiloğlu, 2019). For instance, in Spier-Dance et al.'s (2005) study, the authors investigated and compared four sections of an introductory chemistry course in a college chemistry classroom for prospective science majors. In one section students generated analogies whereas in the other sections students were taught with teacher-generated analogies. The target model corresponded to the accepted and expected version to be understood of the expert consensus model (e.g., "Oxidation involves the loss of electrons", "Size of the outer electron shell typically increases for elements going down a group in the periodic table"; Spier-Dance et al., 2005, p. 167). By applying a two-part question on a final exam, students' responses were scored and coded by the researchers based on students' logical explanation regarding the role of the sizes of the atoms and the differentiation of the attraction for electrons. The results showed that the implementation of student-generated analogies was particularly useful to significantly improve the understanding of those students who often perform below average. Moreover, the analogies generated by students supported the understanding of disciplinary core ideas.

When reasoning with models, the use of analogies has been identified as potentially beneficial to help students understand causal and mechanistic explanations about a system or phenomenon. For instance, Han and Kim (2019) investigated the use of analogy to help elementary students

identify hidden mechanisms involved in the human respiratory system in terms of the *structure* or the components of a system (e.g., lungs and bronchus), *behavior* (e.g., air flows from high pressure to low pressure when the diaphragm moves), and *function* (e.g., during inhalation, air flows from the nose to the lungs). In four classes, students were engaged in modeling tasks in which they elaborated drawing to represent their models, used analogies as biological models (e.g., mechanism of air movement in a syringe), and identified the limitation of the analogies used in class. The analysis of their analogical reasoning showed that the use of analogy allowed students to determine the relationship between phenomenon-based reasoning (inferring how air might pass between the inside and outside the thoracic cavity) and model-based reasoning (manipulating a syringe and using the concept of pressure). Among these studies, the use of analogies, such as the construction of a mechanism-centered model to represent the human respiratory system, helped researchers identify students' misconceptions in their models and facilitated their assessment of student's reasoning about target models.

iv) Technology Can be Used as a Tool to Assess Students' Generated Models

One of the biggest challenges that teachers and researchers face in MBT is being able to access an unknown students' mental model. The use of technology can facilitate the assessment of students' reasoning with models, for example, by assessing how students incorporate variables into their models in their computational representations or by assessing students' drawings or simulations. Twenty-four articles used technology to help elicit details about students' mental models and to assess their reasoning when incorporating new variables in their models (Aksit & Wiebe, 2019; Bielik et al., 2020; Brady et al., 2015; Buckley et al., 2010; Cisterna et al., 2019; Dauer et al., 2019; Dickes et al., 2016; Fretz et al., 2012; Galperin & Raviolo, 2019; Heijenes et

al., 2018; Hsu et al., 2015; Khan, 2008b; Khan, 2011a, 2011b; King et al., 2019; Lally & Forbes, 2019; Louca & Zacharias, 2015; Peel et al., 2019; Pierson et al., 2020; van Jookingen et al., 2019; Vasconcelos & Kim, 2019; Wilkerson-Jerde et al., 2015; Wilkerson et al., 2018; Xiang & Passmore, 2015). For example, Buckley et al. (2010) used *hypermodels* to assess students' ideas in science about transmission genetics and the inheritance of traits. Hypermodels are computer models that include scaffolding, which are developed through problem-solving activities. In this study, students' modeling practices with hypermodels were observed, and students were assessed and provided with immediate feedback during the class through questions, tasks, and explanations. During the construction of computational models, the students were assessed through questions which asked them, for example, to predict offspring and test their predictions, modify parental genotypes, and assess their models of meiosis by determining parental genotypes that resulted in traits appearing to skip a generation. Other studies have also included the manipulation of computational models to investigate students' progress in their models. Khan (2011b) examined over three semesters how an experienced science teacher used a set of computer simulation software called Chemland. The author identified a pattern of teaching strategies that involved i) asking students to analyze a large amount of information on different variables and generating dynamic graphs and animations in the simulation, ii) encouraging students to predict and test assumptions before using their simulations, and iii) asking students to revise their models and rerun tests after adding or removing variables in their models. The use of this simulation followed a generate, evaluate, and modify pattern. For example, the teacher assessed students' understanding by asking them to explain anomalies in temperature of the boiling points of different substances (e. g., methanol and water). The use of computer simulation appeared to be particularly useful to support the activities that were related to evaluate

and modify initial relationships. In another study, Pierson et al. (2020) engaged sixth-grade students in a 9-week ecology unit in which participants used computational models to identify patterns in the dialogic interaction. Computational modeling activities, in which students investigated the dynamics of populations of salamander into a system, were implemented by using StarLogo Nova as an agent-based modeling approach. Moreover, students manipulated variables to observe the emergent behavior of the ecosystem when constructing models. Students also generated codes in the computational environment which were refined and evaluated in order to explain the descriptive power of their models, for example, by modifying variables (e.g., adding worms or detritus). This process of manipulation of the computer codes allowed students to test their refined models and allowed the teacher to assess students' understanding while working with their models. Similarly, Louca et al. (2011) engaged eleven- and twelve-year-old students in modeling-based learning by using a modeling tool called Stagecast Creator that allowed the creation of symbolic models through simulations. During programing, students were asked to assign rules and define objects' behavior to analyze different situations when modeling physical phenomena. The results showed that students' initial models accelerated motion, relative motion and diffusion did not include a causal agent (e.g., how velocity changes in a system), and students' non-causal representations did not include any physical entities such as velocity or acceleration. After using the computer-based environment, all students' models included causal explanations and represented entities in the form of variables. As indicated in the studies above, computational models have the advantage of being able to facilitate the process of testing and revising models and theories. Moreover, students can explore their results immediately, for example, by using dynamic modeling software such as Model-It (Fretz et al., 2002) or SageModeler (Bielik et al., 2020).

In conclusion, four major lines of research were identified within the reviewed studies of strategies implemented to assess students' models (i) the elicitation of models through class discussion; ii) the assessment of modeling practices; iii) the assessment of analogical reasoning; and iv) the assessment of students generated models through technology. Some noteworthy findings are highlighted among studies. A clear trend was identified among some scholars who revealed a cyclical pattern during the process of modeling that involved the generation, evaluation, and modification of initial relationships (e. g., Khan, 2007; 2011b; Nunez-Oviedo & Clement, 2019). It was also revealed that the analysis of students' written explanations and written artifacts (Aksit & Wiebe, 2019; Cheng & Lin, 2015; Gülen, 2020; Han & Kim 2019; Hernández et al., 2015; Hester et al., 2018; Ruppert et al., 2019) can offer researchers and teachers information about how students' reason with a model and how the target model can be enriched, for example, when manipulating variables in a computer-generated model. Finally, it is worth mentioning that the majority of the studies focused on the identification and assessment of students' models, for example, from a formative perspective (e. g., pre and post assessment) that researchers used to measure student' acquisition of the target model (see, for example, Aksit & Wiebe, 2019; Bamberger & Davis, 2013; Bielik et al., 2020; Chang & Chang, 2013; Demir & Nambar, 2019; Demirhan & Sahin. 2019; Dickes et al., 2019; Duncan et al., 2016; Gobert, 2000; Jong et al., 2015; King et al., 2019; Lally & Forbes, 2019; Merritt & Krajcik, 2013; Mierdel & Bogner, 2019; Nelson & Davis, 2012; Peel et al., 2019; Rea-Ramirez et al., 2008; Samarapungavan et al., 2017; Shemwell & Capps, 2019; Vergara-Díaz et al., 2020; Xiang & Passmore, 2015; Zangori & Forbes, 2016). Nevertheless, none of the articles mentioned how summative assessment occurred in the science classroom. In other words, none of the studies explored the summative instruments (e. g., exams, rubrics) that teachers designed and

implemented to assess students' mental models. Whereas the studies identified in the literature review detail different strategies and tools used to identify students' models in the science classroom, previous work has not adequately addressed the role that assessment has played in terms of teachers' pedagogy when teaching with models. According to the review of 86 articles, there remains is a need to characterize each of the specific assessment strategies that science teachers implement in their instruction while interacting with their students to understand how the co-construction of models occurs in the science classroom. For example, only 17 studies out of 86 mentioned how ISTs facilitated the construction of models. Nevertheless, among those studies there was not a clear description of the assessment-based instructional strategies regarding how they used assessment to measure students' reasoning with models. Moreover, regarding the assessment of modeling practices, almost half of the articles (40) detailed the maneuvers that ISTs used to encourage students to generate and use their models; nevertheless, none of these articles mentioned the role that those assessment strategies played on monitoring and tracking students' progress towards achieving a more complex understanding of the learning goals regarding a model nor towards the enrichment of their modeling practices. In this vein, my study attempts to situate the research findings within the context of assessment literacy which has not been explored in any of the articles reviewed.

2.3.2.2 Guiding Question 2: What Strategies have been Implemented to Enrich Teachers' Knowledge of MBT and Assessment Literacy in MBT?

This section reports findings to address the guiding question for this review of literature related to the strategies that researchers implemented to enrich teacher ALMBT. Twenty-eight articles out of 116 included in the review of literature reported several strategies that researchers used to

improve teachers' repertoire concerning how to develop and use assessment when encouraging students to think with models (c. f., Bridle & Yezierski, 2011; Campbell et al., 2019; Guy-Gaytán et al., 2019; Kawasaki & Sandoval, 2020; Pierson & Clark, 2019; Sherwood, 2020). Five main themes were identified among the studies, which are detailed below.

i) *Method Courses can Enhance Prospective Teachers Knowledge of the Role of Models and Modeling in the Science Classroom*

In a series of studies, a group of scholars (c. f., Oh, 2019; Schwarz & Gwekwerere, 2007; Schwarz, 2009) have examined the relevance of supporting pre-service teachers in how to teach and assess students' models. Even though these studies did not investigate in-service teachers' assessment literacy in MBT (the purpose of this study), it is informative to review them to understand the importance of preparing prospective teachers when teaching with models because during the study of such methods courses is when current in-service teachers can be exposed to MBT early on and acquire their first repertoires to implement and assess inquiry instruction. Ten studies, out of 28 articles that included answering this guiding question, supported pre-service teachers in the development of their modeling practice (Carpenter et al., 2019; Günther et al., 2019; Harlow et al., 2013; Jimenez-Liso et al., 2019; Kenyon et al., 2011; Nelson & Davis, 2012; Oh, 2010, 2019; Schwarz, 2009; Schwarz & Gwekwerere, 2007), by challenging pre-service teachers to critique their lesson-planning practices and reflect about their approach to teaching science (Nelson & Davis, 2012; Günther et al., 2019; Kenyon et al., 2011; Schwarz, 2009), and by immersing pre-service teachers in MBT during their practicum (Carpenter et al., 2019; Oh, 2010; Schwarz, 2009). Specifically, these studies taught participants about the foundations of MBT and how to acquire knowledge about this approach of teaching science. For example,

studies focused on their reflection on pre-service lesson plans and assessment strategies revealed that prospective teachers can enrich their ideas about how to teach and assess students' reasoning with models. Kenyon, David, and Hug (2011) used multiple data sources such as interview data, assessments (pre/post-test data), classroom videotapes, and pre-service teachers' artifacts to explore how pre-service teachers implemented MBT. In their study, the authors asked the pre-service teachers to critique lesson plans using an MBT approach as a lens to review their instruction. The researchers analyzed pre- and post-tests at the beginning and end of a method course and after teachers reviewed some ideas in modeling. For example, the author asked participants to analyze narrative vignettes describing classroom scenarios to help preservice teachers visualize and exemplify the role of each strategy in instruction. The comparison of pre-service teachers' reflections revealed that after the course, pre-service teachers were more aware of how to incorporate scientific modeling and place attention on their metamodeling knowledge when developing and analyzing their own or others' lesson plans. Nevertheless, the authors observed that preservice teachers struggled to identify and apply evaluation criteria to promote the revision of models. In another study, Nelson and Davis (2012) found that some pre-service teachers can improve their knowledge about how to assess students' modeling practices. In their study, the researchers enriched pre-service teachers knowledge of MBT by designing and implementing a modeling-based elementary science unit in an elementary science teaching methods course. The researchers analyzed data from homework of thirty-five pre-service elementary teachers and conducted interviews with four preservice teachers in which they were asked to carry out think-aloud evaluations of two elementary student-generated scientific models for the topic of evaporation and condensation, which was used as a pre and post-test. During the study of the unit, pre-service teachers observed anchoring phenomena and created their own

scientific models by developing drawings. They were asked to share the criteria used to evaluate the student models shown in the interview and were asked to explain how they felt about their model evaluation before and after participating in the method course (pre- and post-interview). The analysis of the data showed that pre-service teachers' model evaluation knowledge and criteria varied across participants. For instance, at the beginning of the course, participants focused on evaluating the features and aesthetics of the models (e.g., labels), whereas at the end of the course their ideas were enriched by the addition of new criteria such as sense-making and the role of the explanatory power of the models. Nevertheless, participants did not include in their answers the role of models to make predictions to analyze a variety of phenomena. Similar results have also been obtained in Schwarz and Gwekwerere's study (2007) in which these scholars designed and implemented a guided inquiry and modeling instructional framework (EIMA "Engage-Investigate-Model-Apply") to support pre-service elementary teachers in MBT. During the course, twenty-four participants studied the nature of science, prepared lesson plans, conducted scientific investigations, created and applied models, and developed and taught their lesson plan. The researchers found that pre-service teachers initially used models as resources that are given to students to facilitate their understanding of scientific contents; however, after the course, some participants used models as tools to conduct investigations, answer questions, and represent causal or explanatory components. The researchers identified that participants also enriched their epistemological knowledge of models and modeling. Nonetheless, even though pre-service teachers enriched their knowledge of modeling, they still struggled to include models into their lesson plans and showed limited knowledge about how to assess students' explanations. Similar results were also found in Schwarz's (2009) study which showed that some pre-service teachers only focused on the descriptive aspects of objects rather than promoting

generative questions even after receiving preparation in MBT. Hence, the analysis of these studies showed that method courses are helpful to enrich pre-service teachers' knowledge of the foundations of MBT, nevertheless, it seems that research still lacks efficient interventions that help prospective teachers how to develop assessment instruments and implement specific assessment strategies to guide, for example, the process of model revision.

ii) *Professional Development can Help Teachers to Understand the Foundations of MBT and Enrich their Assessment Strategies*

Ten studies out of 28 articles reported on the implementation of professional development to enrich teachers' assessment strategies concerning how to assess models (Bridle & Yezierski, 2011; Campbell et al., 2019; Guy-Gaytán et al., 2019; Kawasaki & Sandoval, 2020; Khan, 2008b, 2011a; Merritt & Krajcik, 2013; Samarapungavan et al., 2017; Sherwood, 2020; Zangori et al., 2015). Most articles (n = 6) reported relatively short-term professional development events in which in-service science teachers reviewed the foundations of MBT and model-based units (Guy-Gaytán et al., 2019; Kawasaki & Sandoval, 2020; Khan, 2011; Merrit & Krajcik, 2013; Samarapungavan et al., 2017; Zanagori et al., 2015). Professional development (PD) refers to the variety of specialized training to improve teachers' professional knowledge (e. g., workshops, summer courses). For example, in two articles, in-service science teachers reviewed how to use science standards and curriculum materials to engage students in modeling (Kawasaki & Sandoval, 2020; Sherwood., 2020). Sherwood (2020) implemented professional development to enrich 22 secondary science teachers' teaching experiences in the context of the US' Next Generation Science Standards (NGSS). Teachers participated in three days of activities that were held after two weeks each to allow the participants to implement and reflect on the strategies

learned during the PD program. The first day teachers studied an ecology unit and worked as science learners. On the second day, teachers analyzed and evaluated their work as a science learner by deconstructing the components of their practices related to modeling and argumentation. On day 3, teachers analyzed classroom videos that illustrated the activities studied in the prior sessions (e.g., evidence-based argumentation). Similarly, Kawasaki and Sandoval (2020) developed and implemented a professional development program for science teachers in the United States. For three days, fifty-two secondary science teachers participated in a six-hour professional development course in which they reviewed the scientific practices included in the NGSS, with particular emphasis on MBT. The teachers studied anchoring phenomenon, developed models, read about the phenomena (e.g., pressure), revised their models after gathering information, and discussed with their peers their lesson plans and activities for the upcoming school year. Participants were selected and then were observed throughout the school year to investigate how they redesigned their lessons. After the professional development, the authors conducted an interview in which they asked the teachers to provide examples of how they implemented the ideas from the PD. Three class observations were then conducted with each of the seven teachers to identify how the teachers applied their revised lessons. The analysis of class observation showed that teachers often taught science concepts before engaging students in science practices and struggled to engage their students in model creation without providing enough prior scientific knowledge. Also, some teachers fostered the construction of concrete models as a formative assessment tool to evaluate students' conceptual understanding; however, they tended to guide students in the elements and features to be included in their models to represent the function and form of some object or mechanism such as an animal cell. Surprisingly, the authors did not discuss the strategies that teachers used in their pedagogy to

challenge students to revise their models which were related to how teachers assessed knowledge, how they encouraged mistake and how they promoted the elicitation of new ideas.

It is noteworthy that in only one of the studies (Merritt & Krajcik, 2013) did the researchers introduce the scoring rubrics to teachers that they used to score students' models of how a particle model of matter changed after being engaged in a model-based chemistry unit. The teachers participated in a two half-day professional development program for the unit and received online support during the process of teaching students. In this professional development course, teachers were taught about how to create and use models to support students' development of the particle model of matter and reviewed how to use a scoring rubric to assess students' models. The teachers taught 15 lessons that increased in the complexity of modeling practices and helped students progress in their learning performance. Even though the instrument was suggested as a tool that could be used by teachers to track student progress during instruction, in this study, it was not investigated how teachers used the rubric to characterize students' models.

Among the reviewed studies of how PD facilitates the enrichment of science teachers' knowledge of the foundations of MBT and their assessment strategies, the reviewed studies revealed that short-term PD can improve teachers' general knowledge of the foundations in MBT. Nevertheless, the review highlighted issues of concern within the design of the PD since after receiving instruction in MBT the studies reported that teachers were able to enrich their epistemological knowledge of models in science education and how to use models in their instruction; however, they still showed difficulties for teachers to implement assessment

strategies to assess students' models. Moreover, only Khan (2011b) studied the strategies that teacher used to assess students' models. In her study the author explored the strategies that four teachers used to assess students' generation, evaluation and modification of models after attending a PD. The participants were part of a group of 35 science teachers who attended a 3-hour session on MBT in which they examined a chemistry lesson by the implementation of models and modeling practices, studied transcripts of MBT scenarios, and participated in class discussions on modeling. Before starting the class observations, an initial interview revealed that teachers did not follow a systematic approach to MBT. Teachers were observed from 5 to 8 weeks in length and their practice was filmed, and then coded by using a classroom observation rubric which determined the frequency of science teachers' methods of instruction per class according to the modeling practices: generating, evaluating, and modifying models. Her findings showed that the four science teachers engaged students in the generation and evaluation of models, for example, by challenging students to find relationships between variables. The generation of models in each case occurred almost twice as often as the evaluation of models, and the engagement in activities that involved the exploration of the predictive power of models was less common than activities that challenge students to identify the explanatory power of the models. Additionally, ISTs did not engage students in an iterative process of testing and revising the explanatory power of their models and students were not challenged to compare or revise their models to modify and refine them. These results revealed that science teachers even after PD need further assistance in how to assess students' models, particularly when they are engaging students in the evaluation and modification of initial models.

iii) *Implementation of Curriculum in MBT and Co-designing Lesson Plans can Help Teachers Enrich their Practices about how to Teach with Models but They Often Struggle to Assess Students Models*

The extensive review of the literature to answer the guiding question related to the strategies that have been implemented to enrich teachers' knowledge of MBT and assessment literacy in MBT showed that a group of studies focused on the implementation of pre-defined curriculum and co-constructing curriculums (co-designing lessons plans) in MBT that teachers used to teach science. In this review of the literature, it was identified that eight studies out of 28 articles investigated the implementation of pre-defined modeling curriculum originally developed by researchers (Bouwma-Gearhart et al., 2009; Bridle & Yezierski, 2011; Guy-Gaytán et al., 2019; Nunez-Oviedo & Clement, 2019; Pierson & Clark, 2019; Raghavan et al. 1998; Samarapungavan et al., 2017; Zangori et al., 2017). For example, Bouwma-Gearhart et al. (2009) revised and implemented a unit to teach the particular nature of matter (PNM) based on the MUST modelling-based curriculum. Two teachers taught this topic by asking students to construct causal models to demonstrate their predictive power. The authors observed that by implementing the model-based curriculum the teachers were able to engage students in the generation of causal models by working in small groups and challenged them to; collect data about phenomena; examine patterns in the data, construct explanations; share their explanations with others and the teacher; and criticize their own work and their peers' models. Nevertheless, their results surprisingly showed that teachers still struggled to implement strategies that promoted the assessment of models by challenging students to revise their models.

Other studies put special attention in engaging teachers in the co-design, critique and implementation of lesson plans in MBT during a professional development course. Four studies out of 28 articles related to the second guiding question (Becker & Jacobsen, 2019; Thompson et al., 2019; Vo et al., 2019; Zangori et al., 2017) fostered the reflection of teachers' pedagogy to explore how their ideas were enriched during the professional development course. For instance, Becker and Jacobsen (2019) worked with Canadian elementary teachers to enrich their repertoire in MBT. The authors used design-based research in which the researcher and the teacher worked collaboratively to co-design the lessons. To improve teacher's ideas about the use of models in science education, the researcher and the teacher revised and discussed the relevance of developing models to help scientists comprehend complex ideas. The teacher's doubts about the challenges that might be faced in class when engaging students in developing models were discussed with the researcher. Pre- and post-interviews with the teacher and the analysis of lesson plans and video recordings were used to characterize how the teacher assessed students when creating and thinking with their models. In this study, the authors explored how the teacher helped elicit students' ideas about the night sky and facilitated the understanding of mathematical models and orbits to formulate predictions about planets. The teacher's initial reflections about the implementation of MBT showed that the teacher mostly used models to represent the real world instead of engaging students in modeling practices. After revising the foundations of MBT and working collaboratively with the researchers to co-design the lessons, the analysis of interviews and class observations showed that the teacher engaged her students in the formulation of questions and challenged them to elaborate their models. In another study, Thompson et al. (2019) engaged seventh- and eighth-grade teachers in a professional learning community to investigate how teachers shifted their instructional practice with scientific

modeling. Five middle school science teachers co-planned and co-taught lessons focused on scientific modeling. In this study, teachers negotiated the pedagogical resources and reflected on the best way to support students in the generation of scientific models and explanations. During co-teaching sessions, a host teacher led the class, and other teachers and researchers monitored student progress by video recording the class and asking questions to students when thinking with models. Students' written artifacts and interventions in the class were analyzed by the teachers who discussed modifications that they might include in the future. Teachers were observed five times, and their performance was rated based on how they included ideas related to scaffolding modeling in their dialogue.

Two studies also engaged teachers in longitudinal investigations of teachers' knowledge of MBT and enhancing their repertoire of it over the years. These studies (Vo et al., 2019; Zangori et al., 2017) investigated the pedagogy of elementary teachers who had not participated in professional learning experiences in MBT. In Zangori et al.'s (2017) study, the researchers investigated how a modeling-enhanced curricular unit supported third-grade students' explanations about groundwater by thinking with models. The teachers participated in week-long workshops during two consecutive summers. In the first year, participants reviewed the epistemic features of models and their nature and purpose, explored how to engage students in modeling practices, and developed models for the lessons related to the water unit. In the second year, teachers were asked to implement the unit. Also, the researchers asked the teachers at the beginning of the study and after the second year how they enacted a modeling-enhanced unit to support their students' models for groundwater to explore their pedagogical reasoning. Regarding teachers' pedagogical reasoning, the results showed that participants started with a similar level of

knowledge about how to support students in modeling the water cycle; however, in the second year, two of the five teachers were able to offer better support for students' formulation of their models in terms of its components and explanatory power. The results of studies such as the one conducted by Zangori et al. showed that these teachers, who prior to the study had not experienced professional development or any learning experience in MBT, could enrich their repertoire to assess students when they are involved in long-term professional development courses that focus on mastering the foundations of MBT and reflecting on the lesson plans and artifacts used to assess their students. In another study conducted by the same group of scholars, Vo et al. (2019) conducted a longitudinal study of four primary in-service teachers' implementation of MBT with third-grade students relating to hydrological phenomena. These teachers were selected from a group of teachers who had participated in a multi-year professional development course. Participants, who each had ten or more years of experience and no specialization in science education, were involved in a supporting project for elementary teachers over three years to enrich their pedagogy in MBT. Teachers participated in professional development and collaborative work within two summer workshops. The in-service teachers then taught an 8-week unit related to the topic of water and were provided with researcher-developed course materials and lesson plans, including hands-on investigations. In the first year, teachers familiarized themselves with the water unit and reviewed two supplemental pre/post unit lessons and student modeling tasks to help students improve their modeling practices. After the first year, the participants and researchers discussed teachers' ideas regarding how to implement MBT in a one-week professional development workshop on the water unit. Teachers reviewed simulations and a mathematical model used to represent water flow and revised the unit originally developed by the researchers to include modifications in the lessons. The revised

version of the unit was distributed among teachers before the second year, and teachers taught the modeling-enhanced water unit. Another workshop was conducted in summer 2 to enrich teachers' pedagogy in MBT and their content knowledge. Finally, in year 3, teachers reviewed the modeling-enhanced water unit again. Teachers were interviewed five times in the year (spaced 8-weeks apart) and observed (5-6 videos each during the 8-week unit over three years). The results showed that this group of four teachers had a sophisticated knowledge of modeling practices that was used to engage students in the construction of water cycle models. Teachers' practices were enhanced over the years and became more in alignment with the assessment of modeling practices. In the last year of the intervention, participants were able to satisfactorily engage their students in the construction and revision of their models by challenging them to analyze mechanisms and evidence. Nonetheless, it is worth noting that even though teachers' conceptualization for scientific modeling was initially enriched during the PD, their classroom enactment of modeling did not occur simultaneously and was enriched and observed after one year or later.

The results of the ten studies reviewed in this literature review suggest that teachers benefit from ongoing support to learn the foundations of MBT. Interestingly, none of the studies explicitly discussed the impact that method courses or PD had on teachers' assessment literacy in MBT. Overall, it was also identified among the studies that when teachers are provided with model-based curriculum materials and explicit guidelines to implement MBT (c. f., Pierson & Clark, 2019; Raghavan et al., 1998), for example, through PD, teachers are able to show changes in their pedagogy when engaging students in modeling. Nevertheless, the majority of the research showed that many teachers need constant support when choosing and implementing specific

practices to assess students' models. Even though my study does not investigate the impact of an online professional development course on in-service science teachers' pedagogy, this study is intended to address a gap in the literature characterizing science teachers' assessment practices in MBT. My review of literature revealed that assessment literacy in MBT has not been explicitly covered in prior studies. Unlike the studies reviewed in the literature in which teachers were guided in their implementation of curricula while teaching with models, in my study, science teachers were observed implementing the science curriculum without guiding them in the process initially. I made this decision to establish a baseline in ALMBT and identify teachers more often and less often engaged in assessment practices. Finally, I agree that professional development can help teachers understand the foundations of MBT and enrich their assessment strategies. Nevertheless, I also believe science teachers' assessment literacy and the capacity for professional development is partly influenced by their context and is culturally situated (Xu & Brown, 2016). In other words, science teachers' background (e.g., years of teaching experience), context (e.g., type of school, characteristics of students); personal experiences (e. g., pedagogical content knowledge (Park & Oliver, 2008), and prior learning experience in MBT (Zangori et al., 2017) influence the impact that professional development might have on teachers' knowledge base for each of the dimensions in assessment literacy (see Fig. 1).

2.3.2.3 Guiding Question 3: What do the Findings from a Review of Literature Tell us Overall About Science Teachers' Assessment Literacy in MBT?

To answer each of the guiding questions I reviewed 116 empirical articles from 1980 to 2020 related to model-based teaching. In the literature review questions above, I summarized the findings by answering each of the two questions related to i) the strategies that researchers and

teachers have used to assess students' models in the research, and ii) the strategies that have been reported to enrich science teachers' knowledge of MBT and ALMBT. Each of the articles were analyzed to identify assessment strategies that can be used to assess students' models in science. Hence, these articles can inform our understanding of science teachers' ALMBT. In this section, I not only highlight the main findings from the review of literature related to the three previously referenced guiding questions for this review, but I also detail the quantity of studies related to each of the theoretical dimensions used to define ALMBT in this dissertation. It is worth noting that no new articles are presented, and they correspond to a regrouping of the articles based on each of the theoretical dimensions of ALMBT. Moreover, I discuss gaps in the field of ALMBT in this sub-section.

Fifty articles out of the one hundred and sixteen articles analyzed teachers' pedagogy and their assessment strategies to teach and assess students in MBT. In nineteen of the fifty articles examined, the authors explored how teachers' *disciplinary knowledge and PCK* about the foundations of MBT can shape a teacher's assessment strategies to engage and assess students' reasoning with models (c. f., Becker & Jacobsen, 2019; Bouwma-Gearhart et al., 2009; Guy-Gaytán et al., 2019; Kawasaki & Sandoval, 2020; Ke & Schwarz, 2019; Kenyon et al., 2011; Khan 2001, 2008a, 2008b, 2011a, 2011b; Lamar et al., 2018; Nunez-Oviedo & Clement, 2019; Oh, 2010; Tay & Yeo, 2018; Vo et al., 2019; Werner et al., 2019; Wilkerson et al., 2018; Williams & Clement, 2015; Windschitl et al., 2008). For example, in the literature, it was found that analogies can facilitate the elicitation of students' models (Khan, 2007; 2011a; Oh, 2010). Also, science teachers' knowledge of the explanatory and predictive power of models helps them implement strategies to assess students' generated models and their utility (Khan, 2007; 2008a;

Nunez-Oviedo & Clement, 2019; Tay & Yao, 2018; Werner et al., 2019; Williams & Clement, 2015; Windschitl et al., 2008). Three studies out of 19 articles reported that teachers have a restricted knowledge of models and modeling, which might have limited their strategies to assess students modeling practices (Becker & Jacobsen, 2019; Kawasaki & Sandoval, 2020; Vo et al., 2019). Three studies revealed that teachers often convey information when teaching curricular models instead of promoting modeling practices among their students (Guy-Gaytán et al., 2019; Kawasaki & Sandoval, 2020; Wilkerson et al., 2018). Furthermore, some teachers unintentionally guide students to the curricular model instead of using error to help students revise and refine their models (Bouwma-Gearhart et al., 2009; Guy-Gaytán et al., 2019; Kawasaki & Sandoval, 2020; Lamar et al., 2018).

Seven articles discussed the dimension related to the *purpose of assessment, content and methods* that teachers implemented when assessing their students (Ke & Schwarz, 2019; Khan 2008b; Louca & Zacharias, 2015; Mendonça & Justi, 2014; Nunez-Oviedo & Clement, 2019; Pierson et al., 2020; Sung & Oh, 2018). For example, Khan (2008b) found that teachers can assess students' reasoning with a model by engaging students in testing models through the analysis of "what if" scenarios. Also, it was reported that teachers use, for example, formative assessment to help students elicit their models and clarify their own understanding of a model (Ke & Schwarz, 2019; Mendonça & Justi, 2014; Pierson et al., 2020; Sung & Oh, 2018). When teachers implement assessment strategies, they need to be able to interpret the data collected from the assessment. Twenty-two articles described how teachers *interpreted or communicated the results of their assessments* when teaching with models (c. f., Baek & Schwarz, 2015; Bouwma-Gearhart et al., 2009; Campbell et al., 2012, 2019; Cheng & Lin, 2015; Guy-Gaytán et al., 2019;

Hsu et al., 2015; Justi et al., 2009; Kawasaki & Sandoval, 2020; Ke & Schwarz, 2019; Khan, 2008a; Lamar et al., 2018; Louca & Zacharias, 2015; Mendonça & Justi, 2011, 2014; Nunez-Oviedo & Clement, 2019; Oh, 2010; Ryu et al., 2015; Samarapungavan et al., 2017; Sung & Oh, 2018; Tay & Yeo, 2018; Williams & Clement, 2015). Formulating driving questions was a practice that teachers commonly used to judge students' reasoning with models. Driving questions corresponded to specific questions that teachers use to engage students in a process of revision of their initial ideas in order to help them elicit their explanations about a model or phenomenon. This strategy was identified in 15 articles (c. f., Campbell et al., 2012; Campbell et al., 2012; Ke & Schwarz, 2019; Khan, 2008a; Lamar et al., 2018; Louca & Zacharias, 2015; Nunez-Oviedo & Clement, 2020; Oh, 2010; Ryu et al., 2015; Sung & Oh, 2018; Tay & Yeo, 2018; Williams & Clement, 2015). Monitoring students when thinking with models was a strategy found in 8 articles that some teachers used to judge and assess students' reasoning (Baek & Schwarz, 2015; Campbell et al., 2012; Cheng & Lin, 2015; Hsu et al., 2014; Justi et al., 2009; Louca & Zacharias, 2015; Nunez-Oviedo & Clement, 2015; Williams & Clement, 2015). Three articles (Bouwma-Gearhart et al., 2009; Guy-Gaytán et al., 2019; Lamar et al., 2018) reported that some teachers inadvertently do not reflect on the impact of their assessment practices in MBT. Furthermore, Guy-Gaytán et al. (2019) found that some teachers do not guide students during the elaboration and evaluation of their models making it difficult for students to revise and test their models. The same findings have also been identified earlier in Khan's (2011a) study. Hence, effective student guidance during the process of generation, evaluation, and modification of models is crucial to help students progress in their learning (Khan, 2007).

Regarding knowledge of *learning progression and scaffolding*, which corresponded to the new dimension that I included in the definition of ALMBT and showed in Figure 1, nine articles mentioned or used a framework related to learning progression to explore the role that teachers have in scaffolding and helping students to enrich their understanding when working with models (Baek & Schwarz, 2015; Bamberger & Davis 2013; Cheng & Brown, 2015; Ryu et al., 2015; Schwarz, 2009; Schwarz et al., 2009; 2012; Thompson et al., 2019; van Joolingen et al., 2019). These aforementioned studies found that teacher guidance and the generation of scaffolding questions are pivotal to help students progress in the construction and revision of their models, especially when students need to enhance their modeling practices. Based on Figure 1., it can be noticed that teachers also require the ability to *communicate feedback* efficiently to help students evaluate, refine, and enrich their ideas when thinking with models. Only seven articles covered how teachers provided feedback to their students when thinking with models (Guy-Gaytán et al., 2019; Hokayem & Schwarz, 2014; Khan, 2007; 2011a; Mendonça & Justi, 2011; 2013; Nunez-Oviedo & Clement, 2020). For example, Mendonça and Justi (2013) found that when providing background information, teachers can help students revise and enrich their models. Other studies have shown that teachers can promote the generation of consensus models to help students in class achieve a similar understanding of the curricular model to be taught (Khan, 2007; Hokayem & Schwarz, 2014; Mendonça & Justi, 2011; Nunez-Oviedo & Clement, 2020). In Guy-Gaytán et al.'s (2019) study, the researchers found that teachers displayed poor feedback practices, which scarcely promoted the revision and refinement of their models. When teachers communicate feedback based on their assessment repertoires, they need to find strategies to provide feedback not only to the whole class but also to each individual student. Modeling is a process that involves the social construction of students' expressed ideas

that can be enriched collectively. Therefore, and as suggested by the theoretical framework in Figure 1., teachers need to also be aware *of ethical issues in the assessment* (knowledge of assessment ethics), such as confidentiality, in order to provide each student equitable opportunities to participate and grow in class so as to achieve the learning objectives. Only two studies (Guy-Gaytán et al., 2019; Heijnes et al., 2018) explicitly mentioned ethics within professional practices displayed by teachers when assessing their students. For example, Guy-Gaytán et al. (2019) reported that some teachers inadvertently ignored students' answers when they assessed formatively their students. In the case of Heijnes et al. (2018), the researchers found that teachers' assessment strategies were important to motivate reluctant and hesitant students to participate and ask questions during the class. For instance, the researchers pointed out that written assignment can be useful to help students to structure their reasoning processes during modeling; however, the way that teachers provide feedback and support their students during an assessment, for example by scaffolding practices, is essential to facilitate the expression of their models.

In relation to teachers' *knowledge of peer and self-assessment,* I pointed out in my theoretical framework that how teachers offer their students opportunities to reflect on their own learning experience to evaluate their ideas and their assumptions when they develop, revise, and refine their generated models, might be essential to helping students develop modeling practices. In an analysis of classroom teaching and learning progression, I found eighteen articles that detailed how teachers engaged their students in assessment through the process of revision and evaluation of their generated models (c. f., Aliberas et al., 2019; Campbell et al., 2012; Fretz et al., 2012; Gray & Rogan-Klyve, 2018; Guy-Gaytán et al., 2019; Hernández et al., 2015; Kawasaki &

Sandoval, 2020; Khan, 2011a; Lamar et al., 2018; Mendonça & Justi, 2011; Nunez-Oviedo & Clement, 2008, 2019; Pluta et al., 2011; Ryu et al., 2015; Schwarz et al., 2009; Vo et al., 2019; Williams & Clement, 2015; Xiang & Passmore, 2015). For instance, some articles pointed out that teachers have an important role in assisting students in the revision of their models, which can improve students modeling practices, for example, related to model evaluation (Aliberas et al., 2019; Fretz et al., 2002; Ryu et al., 2015; Schwarz 2009). Interestingly, six of the 18 studies analyzed reported that teachers rarely engaged students in assessment when teaching with models (Gray & Rogan-Klyve, 2018; Guy-Gaytán et al., 2019; Kawasaki & Sandoval, 2020; Khan, 2011; Lamar et al., 2018; Vo et al., 2019). For example, Khan (2011a) identified that teachers who implement MBT in their pedagogy often struggled to facilitate the evaluation and modification of students-generated models. These two phases of the GEM cycle are key components that can promote students' engagement in the assessment of their generated models. It is worth pointing out that teachers who receive adequate training in MBT are more able to adopt model-based curriculum. For example, they can encourage students to develop and use epistemic criteria to assess good scientific modeling, for example, in terms of their explanatory and predictive power (Pluta et al., 2011). Nevertheless, based on the studies mentioned above it seems that many teachers need major guidance and support not only regarding how to implement assessment strategies to engage students in the assessment of their models but also in relation to how to assess how students evaluate their own and others' models in the classroom. Finally, regarding the theoretical dimension of *knowledge of grading,* none of the analyzed articles suggested any explicit avenues for the implementation of assessment instruments. Only the study conducted by Merritt and Krajcik (2013) oriented teachers on how to implement a scoring guide

for assessing students' explanation of diffusion of gases model; however, it did not discuss how teachers used the assessment tool.

To sum up, in general, ALMBT has not been widely studied in the MBT literature. Although there are studies that have detailed the role that teachers' pedagogy has on the elicitation and assessment of students' models, there is not enough literature covering each of the dimensions related to ALMBT. The review of literature also revealed that even though some studies explored aspects of teachers' assessment strategies in MBT, these studies did not integrate a comprehensive conceptualization of ALMBT in which each theoretical dimension was carefully detailed and studied. Moreover, many studies have focused on developing professional development courses to enrich teachers' epistemological knowledge of models and modeling in science, but they have not taken into special consideration how to teach teachers to collect evidence from the assessment of students' models and interpret it in order to reshape their pedagogy. Major gaps in the literature are particularly related to the analysis of teachers' artifacts and the scoring tools developed by teachers to assess students' models. None of the reviewed studies investigated this component of ALMBT. Each of the studies detailed above has made valuable contributions to the field of model-based teaching. Nonetheless, there appears to be a further need for both conceptualization and characterization of strategies related to assessment to comprehend how each of the components of a science teachers' knowledge base of assessment (e. g., knowledge of peer and self-assessment; knowledge of grading; knowledge of scaffolding and learning progression) shape how they design, implement, and use assessment when teaching with models in this case. I think that being unaware of the influence of each of these dimensions of ALMBT might limit the potential that science teachers' developed instruments have on their

pedagogy and students learning. Hence, for example, early assessment literate teachers might incompletely use and interpret the results of assessment to shape their pedagogy and offer their students limited opportunities to enrich their models and modeling practices. By knowing about assessment literacy, science teachers might be able to plan, design and implement various types of assessment and gather information from assessment to adjust their instruction when assessing students' reasoning with models. By doing so, I argue that learning can also be enhanced since science teachers can understand how assessment can be used to inform instruction, monitor students' progress, and guide students in the revisions and modification of their models. In this sense, this study attempts to offer for the first time a clear conceptualization of ALMBT to help researchers further their understanding of this construct (ALMBT), in order to be able to help science teachers understand how to be assessment literate in MBT.

Chapter 3: Research Methodology and Methods

In this chapter I introduce the research methodology in which I detail the research paradigm, research setting, research design and research methods. A description of the data collection procedures and data collection instruments used for the development of a baseline of ISTs' assessment literacy in MBT and for the identification and development of the assessment literacy of teachers are detailed below. In particular, I present the methodology and methods used to investigate i) whether in-service science teachers' (ISTs) knowledge of models and modeling was related to their assessment literacy in model-based teaching (ALMBT) and to explore ii) in what ways ISTs' ALMBT influenced their pedagogy.

3.1 Research Methodology

3.1.1 Interpretive Paradigm

This study follows an interpretive paradigm. An *interpretive paradigm* was used to ascertain ISTs' teaching practices about MBT and to explore teachers' assessment literacy in MBT. To understand the philosophical assumptions included in interpretive research, it is important to clarify some concepts related to the research paradigm, such as paradigm, methodology, method, and research design. The notion of paradigm was popularized by the scholar Thomas Kuhn (1962) when he published, *The Structure of Scientific Revolutions*. He suggested that paradigms allow researchers to understand and describe the real world from a specific perspective, and this perspective includes beliefs, values, and methods that are shared by scientists in a particular discipline. McGregor and Murnane (2010) define a paradigm; however, as "[A] set of assumptions, concepts, values, and practices that constitutes a way of viewing reality for the

community that shares them" (p. 419). Broadly speaking, it has been asserted that there are two main paradigms: positivism and post-positivism (Hammersley, 2019). The positivist paradigm is widely used in natural science, whereas in social sciences, post-positivistic paradigms are commonly employed. On the one hand, the *positivistic paradigm* includes the implementation of the scientific method to design and conduct studies and relies on testing hypotheses, experimenting, and observing systems. On the other hand, the *post-positivistic (non-positivist) paradigms* consider the generation of hypotheses by inductive reasoning in order to interpret and understand the real world. In this paradigm, social science researchers often try to understand and explain people's behavior. Participants are studied in their natural settings rather than being part of research in experimental conditions. In the case of post-positivistic paradigms, some examples include critical theory and participatory research (Ponterotto, 2005).

When researchers conduct research, they choose a paradigm to frame their study. Therefore, the *methodology* that researchers use in their study is influenced by the choice of a paradigm. This selection of paradigms requires investigators to make certain philosophical assumptions, which are known as *principles* or *axioms* (McGregor & Murnane, 2010). In this vein, methodology explains the research process and, according to McGregor and Murnane (2010), is constituted by four axioms that include:

- Epistemology: refers to what is considered as knowledge and how the world is studied.
- Ontology: refers to what is considered necessary to be studied and what is considered as part of reality.
- Logic: these are the inferences and assumptions to develop rigorous arguments.

- Axiology: is related to the fundamental values involved in the methodology that researchers implement, for instance, ethics involved in the interaction between the researcher and the participants.

After choosing a paradigm and the methodology, it is important to select the *method* (techniques and procedures) that will help the researcher to conduct the study. For example, in a positivistic paradigm, quantitative research usually includes surveys, field experiments, and quasi-experiments. In a post-positivistic paradigm, methods can include, for example, storytelling, thematic analysis, discourse analysis, action research, critical analysis, and phenomenology. Creswell (2010) states that some studies can navigate among both methods by implementing a mixed methods at the design or methods level in which the connection and integration of quantitative and qualitative data is used to answer a research question.

Regarding research design, Creswell and Clark (2011) define it as the procedures used in a study to collect, analyze, interpret, and report data. In the case of mixed-method studies, these authors identify four main research designs for interpretive research which include; *convergent parallel design* (concurrent qualitative and quantitative data collection with separate analysis); *the explanatory sequential design* (the study starts with the data collection of quantitative data which is then followed by the collection and analysis of qualitative data)*, the exploratory sequential design* (the sequential timing includes the collection and analysis of qualitative data which is followed by quantitative data), and *the embedded design* ("[T]he researcher collects and analyzes both qualitative data and qualitative data within a traditional quantitative or qualitative design"; Creswell & Clark, 2011, p. 71). I use an interpretive paradigm that assumes a naturalistic

methodology by gathering data from different sources such as interviews and participants' discourse (Kivunja & Kuvini, 2017). Interpretivism is a paradigm that relies on the assumption that knowledge is a personal construction that is influenced by participants' and researchers' experiences, and social and cultural backgrounds (McChesney & Aldridge, 2019; Schwandt, 1998; Willis, 2007). This paradigm encompasses inductive reasoning based on the information obtained from participants in their natural setting (Ketokivi & Mantere, 2010). This study sought to explore the experiences of ISTs teaching in middle and secondary schools with the purpose of uncovering how teachers guide their pedagogy and assess students through the personal construction, revision, and evaluation of models. Moreover, interpretive research is underpinned by an ontological relativism that assumes that there are multiple realities that depend on the context in which they occur (Levers, 2013). In other words, individuals' experiences contribute to creating a subjective reality based on people's minds and interactions with others. This research also relies on a constructivist epistemology that assumes that knowledge is constructed by the participants and can be expressed in different ways such as experiences and stories (Ültanır, 2012). Therefore, it is acknowledged that knowledge is subjective and depends on context and culture in this research. Interpretive research was considered appropriate for this study because I assume that science teachers' ALMBT is shaped by their human experiences and social contexts (e. g., type of school, educational background) which influence how teachers assess their students when thinking with models.

3.1.2 Case Study and Cross-Case Analysis

VanWynsberghe and Khan (2007) define case study as a methodology and a "*transparadigmatic and transdisciplinary heuristic that involves the careful delineation of the phenomena for which*

evidence is being collected (event, concept, program, process)" (p. 84). Methodology refers to the "rationale and the philosophical assumptions that underlie any natural, social or human science study" (Campbell, p. 658). The use of case studies attempts to identify and facilitate the construction of a unit of analysis for the phenomenon for which evidence is collected (VanWynsberghe & Khan, 2007) and is characterized by seven common features: (i) *small sample* (N) *for the study*, (ii) *contextual detail*, (iii) *natural settings*, (iv) *boundedness* (temporal and spatial boundary), (v) *working hypotheses and lessons learned* (generated and assessed throughout the course of the study, (vi) *multiple data sources*, and (vii) *extendibility*. Each of these features was included in this study which allows the research to be classified as a case study.

Cross-case analysis was further chosen to explore how individuals interact and experience the world. Cross-case analysis was used as an approach to analyze and contrast participants based on the units of analyses comprised of events, activities, and processes (Khan & VanWynsberghe, 2008). In this case study, the unit of analysis were the pedagogical practices that teachers used to assess students' reasoning with models and each science teacher represented a case. According to Miles and Huberman (1994), cross-case analysis helps the researcher to identify patterns among participants by identifying similar variables and outcome measures. To be classified as cross-case analysis, Yin (2009) suggests that a study needs to include at least two cases that can be considered as parts of independent studies or as parts of a single study. In this study, five ISTs were studied to investigate how ISTs' ALMBT influenced their pedagogy after attending an online training module for professional development in MBT.

3.1.3 Research Setting

3.1.3.1 Participant Sampling

In the phase of baseline of ISTs assessment literacy in MBT the teachers' sample was purposive and corresponded to middle/secondary science teachers across Canada and Chile. This study started with *volunteer sampling* across Canada and Chile. The initial contact with participants began after the approval of the research project by the University Behavioral Research Ethics Board (BREB). British Columbia (BC) school district approvals were obtained for several districts in the metropolitan area of a major city. In Canada, relevant associations were also contacted via email. These associations were requested to share the invitation with their members via their member listservs. In the case of Chile, I used several approved data collection strategies to invite participants and increase the rate of return, including contacting teachers via email and through school principals. In Canada, 43 ISTs answered the questionnaire (see Table 2).

Table 2

Distribution per Province of the Canadian ISTs

Province	Number	%
British Columbia	26	60.47
Ontario	11	25.58
Alberta	4	9.30
Manitoba	1	2.33
New Brunswick	1	2.33

Canadian ISTs included 18 males and 25 females. The average teaching experience was 11.3 years (SD = 8.2), and the number of hours teaching per week fluctuated from 1 to 5 hours (n=1) to over 41 (n = 3) with an average between 16 to 20 and 21 to 25. The average of students per class ranged from 21 to 25 (M = 5.1). Twenty-four ISTs completed an undergraduate degree in science/education, 16 ISTs had a master's in education, and 3 ISTs held a Ph.D. in science (see

Table 4). In the case of Chile, 373 ISTs completed the questionnaire. Table 3 shows the number of participants per region in Chile and Table 4 shows ISTs' highest degree reached. ISTs of each region of the country were surveyed.

Table 3

Distribution per Region of the Chilean ISTs

Region	Number	%
Arica and Parinacota	4	1.0
Tarapacá	6	1.5
Antofagasta	15	3.8
Atacama	3	0.8
Coquimbo	15	3.8
Valparaíso	39	9.8
Santiago	223	56.0
O'Higgins	17	4.3
Maule	10	2.5
Ñuble	2	.5
Biobío	24	6.0
Araucanía	14	3.5
Los Ríos	9	2.3
Los Lagos	9	2.3
Aysén	2	.50
Magallanes	6	1.5

Note: N = 398, this number includes complete and incomplete respondents.

Table 4

Subjects that ISTs Taught in School and ISTs' Highest Degree Reached

a) Subject	Canada	%	Chile	%
Biology	12	27.9	156	39.2
Chemistry	2	4.7	93	23.4
Physics	9	20.9	60	15.1
Two or more specializations in science (e.g., chemistry and biology, physics and biology)	13	30.2	75	18.8
Other (e.g., engineering, environmental science, forensic science)	6	14.0	2	.5
Natural sciences	1	2.3	12	3.0
b) Degree				
Bachelor's in education	0	0	246	61.8
Bachelor's in science and bachelor's in education	24	55.8	20	5.0
Master's in education or in science education	16	37.2	94	23.6
Master's in science	0	0	31	7.8
Ph.D. education	0	0	1	0.3
Ph.D. science	3	7.0	6	1.5

In the Chilean sample, 275 ISTs were females, and 123 were males (see Table 4). Almost 40% of the ISTs taught biology, 23.4% taught chemistry, 15.1% taught physics, and 18.8% taught two or more subjects. The average of teaching experience was 8.57 years (*SD* = 7.1), with an average of hours teaching weekly that fluctuated from 1 to 5 hours (*n* =3) to over 41 (*n* = 74) with an average between 26 to 30 and 31 to 35. The average of students per class was close to 31 to 35 (*M* = 6.9). More than 60% of the ISTs held a bachelor's in education, 5% obtained both a bachelor's in education and science, 23.6% held a master's in education or in science education, 7.7% held a master's in science, only one participant obtained a Ph.D. in science and six participants held a Ph.D. in science.

In the phase of identification and development of assessment literacy proficiency, I employed *convenience sampling* based on a volunteer sample. This type of non-probabilistic sampling is often used in qualitative studies as a strategy to have access to participants who are willing to contribute to a study (Teddlie & Yo, 2007). Convenience sampling is different from purposive sampling because there is no expert judgement to select a representative sample (Battaglia, 2011). Even though a convenience sample might under- or over-represent the population, it is acknowledged that this sampling may allow for faster exploration of a hypothesis or a phenomenon of interest (Battaglia, 2011). Inclusion criteria included ISTs' who were teaching in private or subsidized schools located in the Santiago Metropolitan Region, Chile. This choice was made because school approvals can be easily obtained in these two types of schools. Also, it was requested that ISTs should i) be able to read and understand articles in English from scientific education journals; ii) have earned a bachelor in education; iii) teach in middle or secondary schools (7th – 12th grade); iv) inform and obtain approval from their principals to

allow class observation and the recording of student-teacher interaction. Exclusion criteria included those i) teaching in public schools; ii) authorized to teach by the Chilean Ministry of Education but do not earn a bachelor in education; iii) unable to allow the recording of ISTs' pedagogy. The sample that participated in the identification and development of assessment literacy in MBT phase was comprised of five in-service science teachers. Figure 4 shows the location of the schools (in ovals).

Figure 4

Map of Chile and Santiago Metropolitan Region

Note: Adapted from commons.wikipedia.org under the creative commons CC0 license.

The participants were teachers who were teaching in 4 different municipalities. I assigned pseudonyms to anonymize participants' names and assure confidentiality. James and Samantha

taught at the same school but different disciplines, biology, and chemistry, respectively. Their school was an elite private school located in a residential area that is distant from the traditional urban centers (Fernández, 2009) with a population of more than 1800 students and an average of 21 students per class. Eliana taught chemistry in a Catholic private school located in an upper middle-class area. School population was more than 1400 students and the average of students per class was 29. Lisa was observed in a subsidized school located in a lower-middle class area. Attendant at this school was comprised of approximately 1200 students with an average of 43 students per class. Lisa was observed in a biology class with ninth-grade students. Finally, Gabriel taught biology in a subsidized school located in a middle-class area. School population was almost 500 with an average of students per class of 27. Table 5 summarizes the number of classes that were observed.

Table 5

Summary of ISTs' Classes Observed

Participant	Grade	Class	Number of students	Number of class observation before the OPDC	Number of class observation after the OPDC	Number of pages of observational data transcribed (Classes and interviews)
James	9th	Biology	30 approx.	6	5	180
Samantha	10th	Chemistry	6	5	5	221
Eliana	11th	Chemistry	15 approx.	5	3	173
Lisa	9th	Biology	35 approx.	4	2	96
Gabriel	9th	Biology	35 approx.	3	2	81
						Total: 751 pages

Note: The number of pages of data transcribed from Lisa and Gabriel were considerably lower because the political context affected the school term and many classes were canceled. The topics taught by each ISTs before and after the OPDC are detailed in Table 6.

Table 6

Number of Classes Observed for each Teacher and Topic Taught in the Session

James		
Date	Time	Topic
23-Sep	45 min	Synapse.
24-Sep	90 min	Effects of drugs during synapse.
1-Oct	90 min	Exam revision/ Nervous system/ Introduction to endocrine system.
7 Oct	45 min	Hormones.
8-Oct	90 min	Feedback and mechanism of hormonal action/ Hormonal regulation of glycemia.
14-Oct	45 min	Glycemia.
		OPDC
11-Nov	45 min	Structure of the testicles.
18-Nov	40 min	Seminiferous tubules (Function).
21-Nov	40 min	Structure and functioning of the seminiferous tubules/ Function of the main structures of the male reproductive system.
26-Nov	90 min	Function of the male reproductive system/ Components of the sperm and semen.
2- Dec	40 min	Exam revision.
Samantha		
Date	Time	Topic
26-Sep	90 min	Instructions about how to write an article.
30-Sep	90 min	Accuracy and precision in chemistry.
3-Oct	45 min	Calculate the concentration of chloride in seawater in ppm.
7-Oct	90 min	Formative assessment/ Doubts/ Summarize the unit.
14-Oct	90 min	Introduction concepts: atomic structure.
		OPDC
11-Nov	90 min	Analyze the types of atoms / Evaluate the order of the atoms.
14-Nov	90 min	Historically recognize the order of elements / define periodic properties.
21-Nov	90 min	Analyze the metallic character of the elements in the framework of the periodic table.
25-Nov	90 min	Chemical Reactions of groups IA and VII A. Comments about Test 3/Solve topic handout number 3.
28-Nov	90 min	Solve handout number 3.
Eliana		
Date	Time	Topic
23-Sep	90 min	Identify the average speed in different equations.
27-Sep	90 min	Calculate reaction orders.
30-Sep	90 min	Chemical kinetic laboratory.
1-Oct	45 min	Resolution handout number 7.
3-Oct	90 min	Finish the laboratory assignment.
		OPDC
11-Nov	90 min	Balancing chemical equations (Handout).
18-Nov	90 min	Construction of a diagram to represent the Chemical equilibrium.
22-Nov	90 min	Variation of K with temperature.

Table 6 (continued)

Lisa		
Date	Time	Topic
9-Oct	70 min	Identify key points of the independent phase of light.
	40 min	Compare the stages involve in the photosynthesis.
	70 min	Identify key points of the cellular respiration.
	40 min	Identify the stages of cellular respiration.
		OPDC
	60 min	Interpret data and evidence of evolution to present arguments about biodiversity.
	60 min	Recognize the importance of the evidence provided by the theory of natural selection by the scientific community.
Gabriel		
Date	Time	Topic
3-Oct	45 min	Explain how the biodiversity was originated.
4-Oct	45 min	Answer questionnaire to summarize unit 3.
10-Oct	45 min	Characteristics of apes.
		OPDC
21-Nov	40 min	Homologoues and analogoues organs.
28-Nov	40 min	Evolucionism.

Note: Due to each school context, it was not possible to observe the same number of classes for each participant; however, methodological triangulation through different sources of data was used to complement the limited number of class observations after the OPDC in the case of Eliana, Lisa and Gabriel.

The following sections present how the data analysis was conducted in the phase of identification and development of ALMBT. Specifically, the data collection timeline, the rationale of the Online Professional Development Course in MBT (OPDC), the data sources, research methods, and issues related to mix-method studies such as trustworthiness are discussed.

3.1.4 Data Collection Timeline

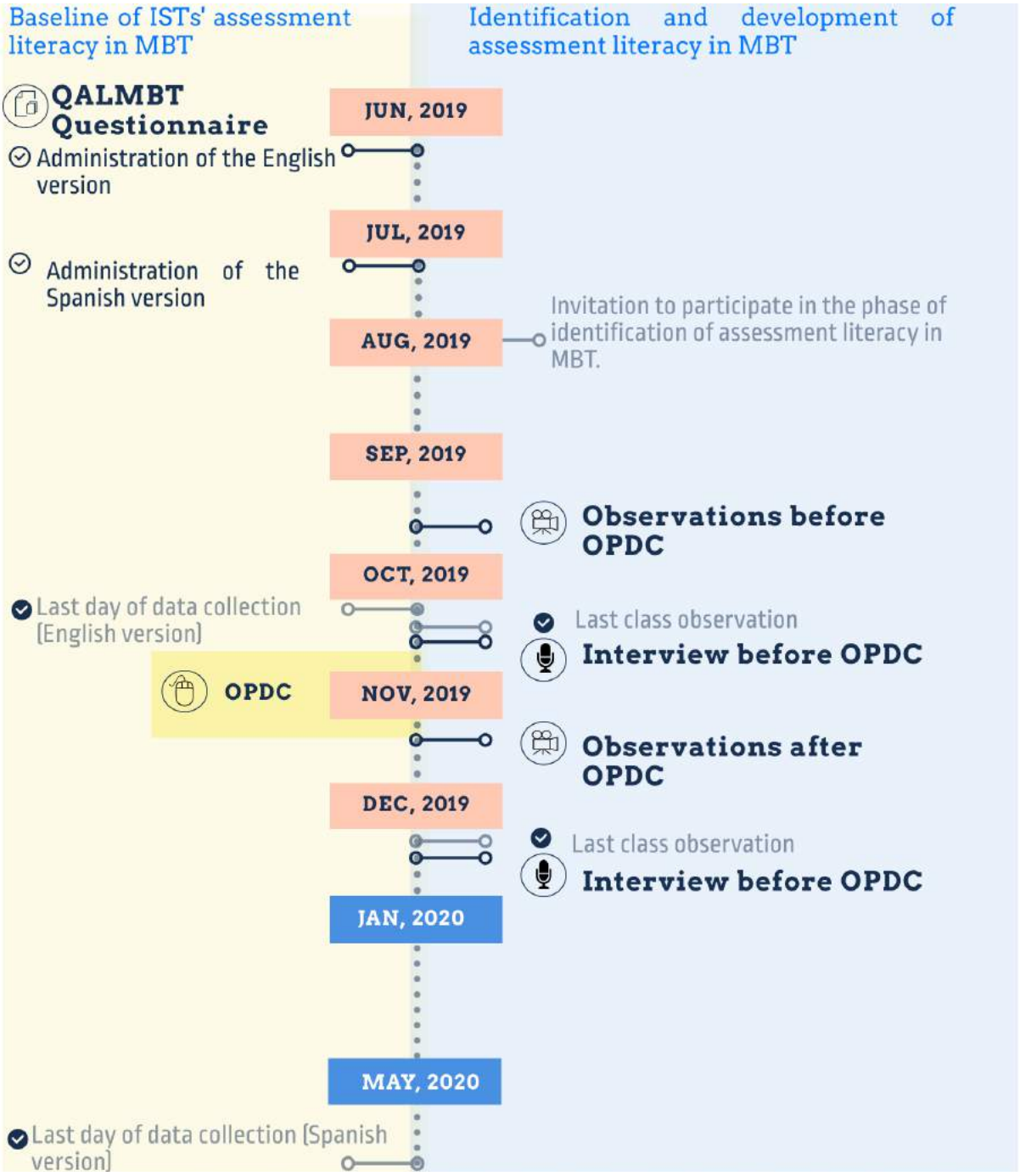
In the first phase, called, baseline of ISTs' assessment literacy in MBT, the data collection started in June 2019 with the questionnaire on assessment in literacy in MBT (QALMBT). ISTs across Canada received the link with the questionnaire first. Once the QALMBT was translated into Spanish, Chilean ISTs received a link with the Spanish version of the questionnaire in July 2019. The English version of the QALMBT questionnaire was administered in early June until early October 2019. The Spanish version of the QALMBT questionnaire was available until

early June 2020. It is worth mentioning that the timelines for the English and Spanish version of the questionnaire were different since I was not able to reach more participants in the Canadian context.

The phase of identification and development of assessment literacy in MBT only was investigated with Chilean ISTs and started in mid-September, 2019. In early August 2019, those ISTs who had already answered the QALMBT questionnaire were invited to participate in the second phase of the study related to the identification of ISTs' assessment literacy in MBT. Only Chilean ISTs from subsidized and private schools from the metropolitan region received this invitation to participate in; class observation before and after the OPDC; interviews before and after the OPDC; and received the invitation to attend an online professional development course in MBT. Six ISTs confirmed their participation in the study; however, one of them withdrew from the study one week before the start of collecting data in schools. I arrived on the field site in Chile on September 23rd, 2019. Class observations took place in the last week of September, and the observations started according to ISTs' availability. Interviews before the OPDC were conducted at the end of the second week of October. Class observations after the OPDC were conducted in the second week of November until the end of the month. In December I was not able to continue observing classes due to early school closures. Due to the social and political crisis of Chilean society which started in mid-October (Garcés, 2019), many classes were canceled, and students finished the school year earlier than usual. Interviews after the OPDC were thus conducted in mid-December 2019. The following diagram (Figure 5) summarizes the main phases of the study.

Figure 5

Data Collection Timeline for Each Phase of the Study



3.1.4.1 Online Professional Development Course (OPDC)

I developed the online course, OPDC, for ISTs in order to inform them about the foundations of MBT. The OPDC was not informed by the results from the questionnaire. In other words, it was designed and planned prior to research. The OPDC was implemented based on different pedagogical materials collected from three main resources i) internet-based activities, ii) excerpts from articles from journals in science education, and iii) my experience as a chemistry and biology teacher. The course was expected to be completed in 10 hours and was developed as an attempt to enrich ISTs' epistemological knowledge of models and modeling. The OPDC was organized based on the following sequence in which each topic was a module: i) how science proceeds and types of reasoning in science; ii) the role of models and students' mental models; iii) modeling practices to facilitate students' understanding in science; iv) instructional strategies in MBT, v) teaching science using models and the process of modeling to implement science curricula; vi) suggestions for assessment in MBT and viii) learning progressions in MBT. This sequence in the content of the OPDC is relevant because it constitutes a progression of topics related to MBT from my perspective. In the OPDC, MBT was presented as a theoretical approach to teaching in which teachers need to engage students in the generation, evaluation, and critique of their models. Models were defined as i) idealized abstractions of ideas, systems, processes, or phenomena that are often used to understand theories (Develaki, 2016; Gilbert, 2004; Passmore et al., 2014); ii) representations used to provide explanations or describe causal relations of the target system (Giere, 2004; Schwarz et al., 2009); iii) simplifications of the attributes and characteristics of a target system (Coll & Lajium, 2011), and iv) sources of knowledge used to generate and test hypothesis (Rupert et al., 2017; Svoboda & Passmore, 2013). In the case of modeling for educational purposes, the OPDC conceptualized modeling as

the process of building critiquing, and modifying students expressed mental models in which science teachers guide students in the refinement of their ideas, abstract concepts or models. For each module, the online course summarized and covered the main ideas for each topic indicated above. Each course page included links to videos on the Internet about models, modeling practices, definitions (e.g., theory, facts, etc.), and also questions that invited teachers to reflect about their teaching. In each module, ISTs were asked to read selected pages of articles in science education to enrich their understanding of the foundations of MBT and modeling. The following paragraphs discuss the content covered in each module in brief.

In the first module, ISTs briefly reviewed how science proceeds and studied different types of reasoning in science. This module provided teachers with a general idea about the nature of science and ensure they know what scientists mean when they use words like theory, hypothesis, law, evidence, and claim. The second module was related to understanding the nature and purpose of models in science education. Definitions of models were included such as a model is an idealization and a conceptual representation (Develaki, 2016) that represents, explains and predicts phenomena or events. In the third module, ISTs reviewed a set of knowledge and skills that students need for the generation, testing, evaluation, refinement, and modification of models (Windschitl et al., 2008). This module particularly focused on the GEM cycle and breaking down the inquiry skills that students need to develop in each stage of the cycle, for example, gathering evidence, elaborating hypothesis, and formulating scientific explanations (Gobert et al., 2011; Khan, 2007). The fourth module examined examples of model-related teaching activities for the science classroom. Examples of case vignettes in the context of MBT classroom activities were provided to help ISTs recognize and interpret data from empirical research. The fifth module

introduced excerpts coming from the Next Generation Science Standards (NGSS) from the United States and from the Chilean science curriculum related to MBT to provide ISTs with examples of curriculum materials that are coordinated with learning activities. For the sixth module, ISTs reviewed how to assess models and scientific modeling tasks, for example, by utilizing rubrics (e.g., criteria to identify components, relationships, and the connection of students' models with a phenomenon; Mayer & Krajcik, 2015). Moreover, ISTs were asked to analyze other instruments and strategies to elicit and assess students' models such as think-aloud protocols (Jonassen & Cho, 2008), the development of artifacts and drawings (Liu & Lin, 2015; Luxford & Bretz, 2013; Quillin & Thomas, 2015) and the use of paper-and-pencil tests (Devetak et al., 2009; Sunyono et al., 2015). Also, this module reviewed works related to the development of epistemic criteria which are useful to help students assess others' models (see, for example, Chin & Brown, 2000; Gobert et al., 2011) and included examples suggested by Krell et al. (2014) to identify students' levels of understanding models and modeling. Finally, in the last module, ISTs studied examples of learning progressions in MBT. This module included examples of learning progression to guide teachers in their pedagogy in order to immerse students in the use of models as generative tools to explain phenomena (Schwarz et al., 2009). In the last class observation before the OPDC, ISTs received a USB flash drive with the course materials (handouts and readings). Each module was in a Word format and included access to the activities which were hosted in the UBC tool provided by Qualtrics. This platform was used as a virtual learning environment, and participants had access to upload documents, submit their answers. It is worth mentioning that data from the course were not collected because teachers prioritize the study of the modules, and they informed me they did not have enough time to complete the suggested activities included in each module.

3.1.4.2 Rationale of the OPDC

The OPDC follows a sequence that considers the nature of science and the foundations of MBT. Professional development is relevant for in-service teachers because it contributes to enrich or modify teachers' cognition, beliefs, and practices, for example, regarding their curricular knowledge (Avalos, 2011). In this vein, the OPDC was based on a professional training course that attempted to inform and expand ISTs' professional expertise about different strategies and skills that they could develop, use, apply and adapt to their classroom (Asghar & Ahmad, 2014) in MBT. I decided to create an OPDC because other strategies such as workshops and a face-to-face professional development course would require attendance at specific times. Such approaches would be difficult for the IST's participating in the study because they worked full time and did not have enough time to participate in extracurricular activities. Moreover, by participating in the OPDC, ISTs could check the course materials according to their availability. It is worth mentioning that the online course was only 10 hours in length, and it is possible that ISTs did not have enough time to dedicate to the activities. Desimone (2009) suggests that professional development interventions require at least 20 hours or more to have a real impact on pedagogical knowledge of teachers. It is recognized that the short amount of time and online nature of the OPDC might have a smaller impact on participants' beliefs and insight about how to teach in MBT compared to longer professional development. The focus of the research was not on evaluating the effectiveness of the OPDC per se. Rather, I attempted to identify how ISTs, based on their own teaching experiences, implemented MBT after attending a 10-hour course in MBT.

The OPDC content and tasks followed traditional professional development activities (Bayar, 2014) (e.g., reading and analyzing articles) that required ISTs' self-regulation of learning. The activities in the modules attempted to provide authentic learning contexts (Vonderwell & Turner, 2005) that helped ISTs analyze examples that might be pertinent to their reality and experiences in the educational system. For the pedagogical framework for the OPDC, I followed three approaches. Firstly, I considered an outcome-based education (OBE) approach to organize and structure the course. The OBE guided the OPDC according to the outcomes that I desired to observe in the participants and the intended learning outcomes (Pang, Ho & Man, 2009). In this study, the learning outcomes were to foster teaching strategies that ISTs might use in their pedagogy to assess students. Secondly, I also adapted a data-based teacher development approach (Borg, 1998) to guide the development of course materials. This approach is similar to the data-based decision making defined by Prenger and Schildkamp (2018) where teachers review and analyze their own or others educational practices to improve and adapt their instruction. The analysis of teaching practices used in my OPDC approach might help teachers to identify the rationale for their pedagogy and react to the vignettes based on their knowledge and experience, "[M]oving from an analysis of another teacher's work to more self-oriented inquiry" (Borg, 1998, p. 279). I designed this teacher development course as a complementary resource that attempts to encourage ISTs to foresee the benefits and limitations of MBT in the classroom and self-evaluate how they might immerse students in the generation, evaluation, and modification of models.

3.1.5 Explanatory Sequential Research Design

My study used an explanatory sequential research design (Creswell & Clark, 2011). In this type of research, "[Q]ualitative (text) data are collected and analyzed second in the sequence and help explain, or elaborate on, the quantitative results obtained in the first phase." (Ivankova, Creswell & Stick, 2006, p. 5). This research design was chosen to follow up the quantitative results (baseline of ISTs' assessment literacy in MBT) with the analysis of qualitative data (in what ways ISTs' assessment literacy about models and modeling influence their pedagogy). I chose an explanatory sequential design because I attempted to use quantitative data to obtain a general idea of ISTs' assessment literacy in MBT. Then, I collected qualitative data to refine and extend the findings from the analysis of the questionnaire instead of collecting additional quantitative data to explain the relationships found in the second phase of the research (exploratory design). In my study, the questionnaire QALMBT was informed by the literature in model-based teaching and assessment literacy.

3.1.6 Data Sources

The data sources in the study included i) the responses to the questionnaire of assessment literacy in MBT (QALMBT), ii) class observation rubrics (Rubric of Assessment Strategies in Models and Modeling "R-ASMM"), and iii) ISTs' artifacts (e. g., exams, lessons plan). Each of the data sources are detailed below.

3.1.6.1 QALMBT Questionnaire

The QALMBT questionnaire was designed as a self-report inventory that asked science teachers how often they implemented specific assessment practices when assessing students of models

and was also used to ask them how much they agreed or disagreed with a particular statement related to the epistemology of models. The administration of the questionnaire was online, and this mode made the access easier for ISTs. Participants could answer the questionnaire from their personal computers at any time. Secondly, to avoid missing data, I conditioned that each question included in the questionnaire must be answered. Thirdly, a self-administered questionnaire was preferable to an interviewer-administered mode in order to reduce a social desirability bias due to the presence of the researcher. One limitation of the administration of the QALMBT online is the fact that the items cannot be clarified; however, participants had access to a blank space at the end of the questionnaire to suggest comments. It is also worth mentioning that the questionnaire completion time was checked for each participant since I anticipated that ISTs would need between 5 and 15 minutes to answer the whole questionnaire. For those ISTs who answered in less than 5 minutes, their answers were re-checked to analyze if they showed variations in their answers for each item. I used questionnaire competition time as a first indicator of questionnaire data quality and check that ISTs carefully answered each item.

3.1.6.2 Development of the Questionnaire

The main goal of the questionnaire was to identify teachers' pedagogy related to teaching about and with models. The questionnaire included three main sections. The first section of the questionnaire asked demographic information (e. g., years of teaching experience, discipline, age). The second section included question related to how often teachers assessed their students when thinking with models, and finally, the third section asked question related to science teachers' epistemological knowledge of models. The process of constructing items for the second and third section of the QALMBT included three steps: a) a review of articles about Likert-type

instruments to measure science teachers' knowledge about models and modeling, b) the study of Likert-type instruments to measure teachers' assessment literacy, and c) the grouping of the items elaborated for the questionnaire according to the dimensions from Figure 1. The grouping included two main categories: assessment literacy in model-based teaching and the nature and purpose of models. The assessment literacy dimensions from Figure 1 were eight (e. g., disciplinary knowledge and PCK, knowledge of feedback, knowledge of grading). It is worth mentioning that it was necessary to include each item stem in the questionnaire twice in order to identify ISTs' assessment literacy in general versus assessment literacy specifically in MBT. Each item included a common statement (stem) that was used to ask the same question twice to avoid duplicating the number of questions. Hence, each common statement is followed by a first sentence that focuses on ISTs' practice when teaching science without emphasizing a particular teaching approach, and a second statement which focuses on ISTs' practice specifically when teaching science by including models. For example, for question 8, the common statement stated, "When I develop summative assessment, I inform students in advance about the criteria that I will use to assess...". Below this statement, two options were indicated. The first option was more general, "... understanding in class.", whereas the second option emphasized on MBT and stated "…their models". In both cases, teachers were asked to indicate how often they included it in their pedagogy. The double statement in each question was included with the objective of avoiding respondent fatigue (Ben-Nun, 2008). To make the analysis of the questionnaire based on ISTs' responses for each question easier, I will refer to the results from the questionnaire as "QALMBT-Generic" for the results obtained for the item in which they did not emphasize a particular teaching approach. The acronym "QALMBT-Modeling" will be used to refer to the results obtained from ISTs' answers based on an MBT approach. After validation by a group of

experts, the number of items in the questionnaire was reduced to 35. It is worth mentioning that science teachers' answers from the questionnaire were particularly valuable to answer research question 1 regarding whether ISTs' knowledge of models and modeling related to ISTs' assessment literacy in MBT, as it is indicated in the following paragraph. Moreover, the results from the second section of the questionnaire were particularly valuable to partially respond to the second research question regarding in what ways ISTs' assessment literacy about models and modeling influence their pedagogy. To do so, both the type of assessment strategy and the frequency of how often the strategy was enacted were also analyzed in ISTs' pedagogy during the classroom observations. The distribution of the items is detailed in Table 7. The number of the item corresponds to the distribution of the item in the questionnaire.

Table 7

Distribution of Items Included in the QALMBT-Generic/Modeling

Dimension	Items				
Knowledge of disciplinary knowledge and PCK	1	6	20	26	
Knowledge of assessment purposes, content and methods	2	10	15	27	31
Knowledge of grading	5	23	28	32	
Knowledge of feedback	3	16	19	33	
Knowledge of assessment interpretation and communication	7	21	24	34	
Knowledge of peer and self-assessment	4	11	13	17	29
Knowledge of assessment ethics	8	22	30	35	
Knowledge of Scaffolding and learning progression	9	12	14	18	25

Note: The items are sorted by dimensions included in the framework of assessment literacy (see Figure 1). The numbers reflect the order in which the item was included in the final version of the questionnaire QALMBT.

Finally, the third section of the questionnaire focused on identifying ISTs' understanding of models (nature, purpose, etc.). I refer to this section of the questionnaire as "QALMBT-Epistemic". In this section, each statement was constructed on teaching about the nature of models. The studies conducted by Grünkorn et al. (2014), Krell et al. (2015), and Gogolin and Krüger (2018) were used as guidelines to design items to be included in this section of the

questionnaire. The number of items related to the epistemology of models and modeling that were included in the QALMBT-Epistemic are detailed in Table 8. The full version of the entire questionnaire is included in appendix B. The number indicates in Table 8 the distribution of the items in the questionnaire. The process of construction of items for this section of the questionnaire is detailed in the following paragraph.

Table 8

Distribution of Items in the QALMBT-Epistemic

Dimension	Items			
Nature of models	1	6	13	18
Multiple models	7	14	19	20
Purpose of models	4	8	15	16
Testing models	2	10	12	17
Changing models	3	5	9	11

Note: The items are sorted by dimensions based on the epistemology of models.

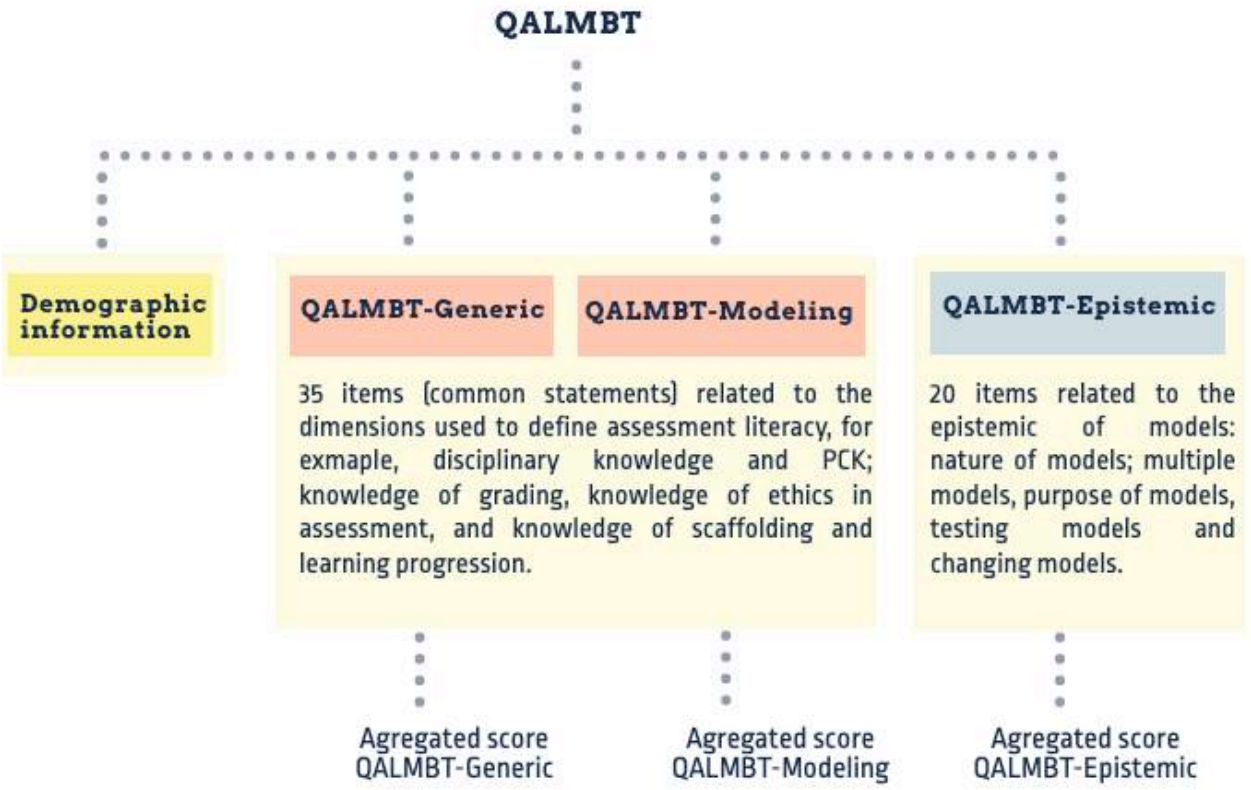
Before the administration of the questionnaire, the studies indicated above were reviewed and used as guidelines for the construction of the items for the third section of the questionnaire (QALMBT-Epistemic). These studies were based on a theoretical framework originally elaborated by Upmeier zu Belzen and Krüger (2010), where three levels of understanding of models as both products and methods of science are suggested. In Grünkorn et al.'s (2014) study, these authors developed and tested a category system of students' understanding of the nature of models (epistemology) that included five aspects: the nature of models, multiple models, the purpose of models, testing models, and changing models. These aspects were organized according to the three levels of complexity Upmeier zu Belzen and Krüger's (2010). For example, in studies regarding students' understanding of the nature of models, these authors identified that the lowest level of complexity (level I) corresponds to the understanding of

models as a replication of the original, level II is related to an idealized representation of the original, and level III (the highest level of understanding) as a theoretical reconstruction of the original. In the QALMBT questionnaire, the items in this section that related to the nature and purpose of models were elaborated upon based on the highest levels of complexity suggested by the authors. The decision to include only the highest level (level III) or the most accurate descriptions of understanding of the nature and purpose of models for teachers was made to reduce the number of items for the questionnaire (extension of the questionnaire) and to be able to cover each dimension regarding the epistemology of models and modeling. For example, for the same dimension (nature of models), in the QALMBT one of the items included "a model is a reduced or theoretical reconstructed part of reality", which in Grünkorn et al.'s (2014) study would correspond to the highest level of understanding. Another reason why only the highest levels of understanding in models and modeling were included in the QALMBT, is the fact that an aggregate score was calculated based on ISTs' responses by using a 5-point Likert scale (1 = strongly disagree and 5 = strongly agree). For example, the maximum aggregated score for the QALMBT-Epistemic was 100 (20 items times 5). In other words, each item was multiplied by the number that each IST selected regarding the extent they agreed to the statement. For example, if an IST selected that s/he agreed that "1. A model is a theoretical construction of reality", the score was 4 for this item. The same procedure was repeated for each item and then an aggregated score was calculated based on the sum of the scores for each item. The aggregate score was calculated to identify a correlation between it and the results obtained from the QALMBT-generic and QALMBT-modeling. This correlation permitted me to answer research question 1 determining if ISTs' knowledge of models and modeling is related to ISTs' ALMBT.

It is also worth mentioning that Likert-scale intervals in each section of the questionnaire were not weighted since I used an ordinal measure (e. g., strongly disagree, disagree) in which I assume that the distance between scale values are described as an ordinal distance instead of being definite intervals.

Figure 6

Structure of the QALMBT Questionnaire



Finally, Figure 6 summarizes the structure of the QALMBT. It should be noted that the QALMBT is comprised of three main sections in which the scores from the QALMBT-Generic and QALMBT-Modeling corresponded to the dependent variable. The aggregated score computed from the QALMBT-Epistemic corresponded to the independent variable. These

variables were used to determine if ISTs' understanding of models and modeling is related to IST's assessment literacy in MBT.

3.1.6.3 Spanish Translation of the QALMBT

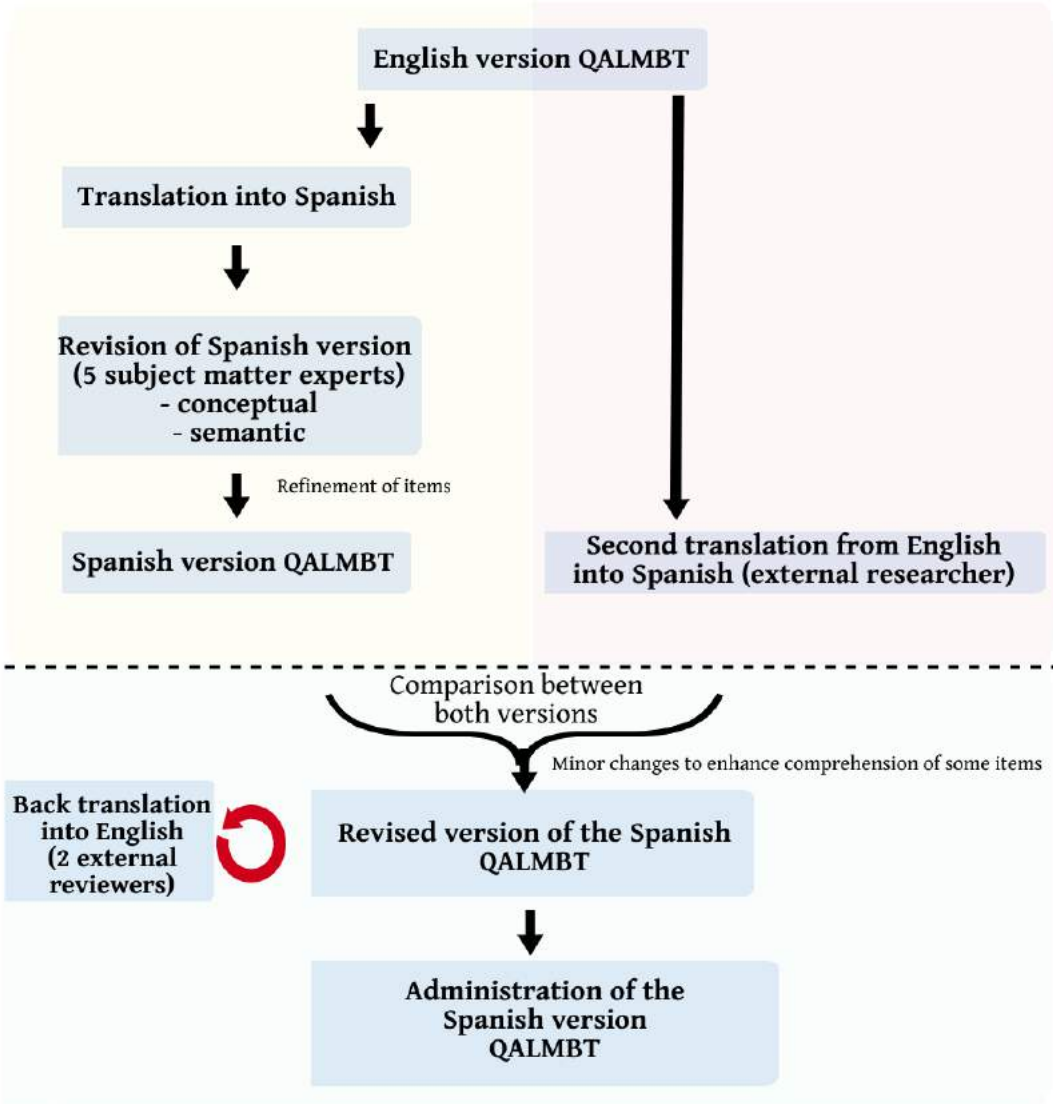
The QALMBT was initially written in English and then translated into Spanish in order to use the same instrument in Chile as well. The purpose of expanding the study to Chile was to survey enough participants to support the construct validity (EFA) of the QALMBT. In factor analysis, at least 100 participants are suggested to conduct this type of analysis well (Mundfrom et al., 2005); however, the Canadian sample was not large enough to validate the instrument used to investigate ISTs' assessment literacy in MBT.

I conducted a direct translation of the questionnaire choosing to do semantic and conceptual equivalence rather than a literal translation. Semantic equivalence refers to maintaining the equivalence of meaning between the original version of the questionnaire and the translated version (Herdman, Fox-Rushby & Badia, 1998). Conceptual equivalence is related to the questionnaire' ability to measure the same construct across different cultures (Flaherty et al., 1988) and the degree that the versions of the questionnaire have to the same relationship of the underlying concept, even though they are administered in two different contexts (Hall et al., 2018; Herdman et al., 1998). In the QALMBT, even though many of the items followed a literal translation since the conceptual equivalence was the same in both languages, for some of the items, I attempted to consider the closest possible meaning to keep the same idea as the original (Hall et al., 2018). A back translation of the QALMBT was conducted to compare the English and Spanish versions of the questionnaire (Behr, 2017). Figure 7 summarizes the stages used to

ensure semantic and conceptual equivalence of the English and Spanish versions of the questionnaire.

Figure 7

Stages Included in the Translation Process to Ensure Semantic and Conceptual Equivalence of the QALMBT



Back translation refers to the translation of a questionnaire back to the original language and the comparison of both versions to identify discrepancies or errors in the translation (Behr, 2016). The back translation from Spanish to English was conducted by two people who were Spanish speakers and fluent in English but not necessarily experts in science education. Taber (2018a) suggests that back translations in science education must report how translation was checked in order to facilitate readers who are not able to read in another language. Five steps are usually suggested to ensure cross-cultural equivalence in the translation of self-report measures. These steps include: i) translation of the instrument by two translators; ii) synthesis and revision of the translation in order to identify discrepancies; iii) back translation into the original language to check for errors in the translation; iv) subject-matter expert revision of the back translation, and v) pretesting of the instrument (Beaton et al., 2000; Quigley et al., 2010). Due to the lack of funding for this portion of the study, the process of back translation of the QALBMT included some modifications detailed below.

Before conducting the back translation of the QALMBT, the translated version of the questionnaire was revised independently by five Spanish native speakers from Chile who had some training or knowledge in research in science education. The Spanish speaking reviewers were a Ph.D. in science education with a bachelor in education in Physics, two science teachers with a master in education, a science teacher with a master in Chemistry, and a science teacher. Each member provided feedback regarding wording and content using the Spanish version of the questionnaire. None of the reviewers identified any problems with the statements included in the questionnaire, indicating that the questionnaire might be understood by science teachers in the Chilean context. The first step of the back translation was conducted by a Ph.D. in science

education who was a Spanish native speaker and fluent in English. This scholar translated the English version of the questionnaire into Spanish. The Spanish version of the questionnaire that I initially translated was compared with the Spanish version suggested by this scholar in order to analyze the adequacy of the modifications in the first version in terms of semantic and conceptual meaning. Discrepancies were discussed until reaching an agreement regarding the conceptual equivalence between the English version and the Spanish version that I initially developed. It is worth mentioning that some items were reworded instead of eliminating them to keep the same psychometric properties of the instrument (Flaherty et al., 1988) between the English and Spanish version. Several examples of rewording are discussed later, and Appendix C presents an example of the back translation of each item. After reaching agreement and discussing each item with another reviewer who was fluent in Spanish and English (see Appendix D), the instrument was also administered in the Chilean context in Spanish. Some examples of rewording are indicated in the next paragraph.

Regarding the QALMBT-Generic and -Modeling, in item 9, I preferred not to include the concept of scaffolding in the translated version of the questionnaire because the literal translation "andamiaje" (scaffolding) did not sound intuitive or natural in the Spanish language. It is worth noting that the word scaffolding might have been encountered by some Chilean science teachers when they studied concepts related to the zone of proximal development and scaffolding suggested by Vygotsky (1978); however, in the Educational Chilean context, many professionals are authorized or certified by the Ministry of Education to teach in schools for a period of time even though they have not earned a Bachelor in Education (OECD, 2003). This is not a usual route to begin a teaching career, but in science education in the Chilean context it has become a

common practice due to the lack of ISTs. Therefore, the word "andamiaje" (scaffolding) might have been unfamiliar for them since professionals who are, for example, biologists or engineers might not use the term nor understand the meaning of the word in an educational context. Hence, the word "scaffolded assignments" in the statement "I design scaffolded assignments or tasks that progress in complexity in order to assess students' understanding about..." was omitted without losing the original meaning as suggested in the following reworded statement "I design tasks that progress in complexity in order to evaluate the understanding of students' about… (Diseño tareas que progresan en complejidad con la finalidad de evaluar la comprensión de los estudiantes sobre…)" in the Spanish version. This decision was reviewed and discussed with the external researcher who also translated the questionnaire into Spanish and who agreed to the modification of this item. In the QALMBT-Epistemic version, minor rewording was needed. For example, in item 3, I changed the word "compels" to "induces" ("inducir" in Spanish) because the translation for the original word is related to "obligue/force" and has a meaning that seems to be always true or suggests a causal relationship. Finally, it is worth mentioning that no items of the original subscale were eliminated after the translation process to cover each of the construct's domains for assessment literacy.

3.1.6.4 Observation Rubric of Assessment Strategies in Models and Modeling (R-ASMM) and Transcriptions of the Lessons

Classroom observations can help teachers and researchers characterize the instructional techniques that teachers implement when interacting with their students in the classroom. In my study, I conducted classroom observations before and after science teachers attended an online professional development course (OPDC) in MBT. I used digital camera recordings and digital

photographs, and the implementation of a classroom observation rubric to register observations of teachers' pedagogy. To characterize ISTs' teaching strategies to assess students in MBT, I developed a rubric for class observations to track the type and frequency of activities that ISTs implemented in the classroom. This rubric is called "Assessment Strategies in Models and Modeling", R-ASMM, and the rubric indicators were elaborated based on studies in MBT and studies in science education (e.g., Bennet, 2017; Furtak et al., 2012; Khan, 2007, 2011b; Pluta et al., 2011). To construct the indicators, scientific modeling was understood as a cyclic process that involves the modeler's prior knowledge and experience (Giere et al., 2006). The R-ASMM included a group of indicators (codes) which were organized based on the definition of assessment literacy used in this study (e.g., disciplinary knowledge and PCK, Knowledge of purpose of assessment, Knowledge of grading) and informed based on other established classroom observation rubrics on the GEM cycle (see, for example, Khan, 2007; 2011b). Examples of indicators are included in Table 9.

Table 9

Examples of Indicators Included in the R-ASMM Classroom Observation Rubric

Theoretical Dimension (Knowledge of)	Code	Description of Teacher Action Observed
Disciplinary Knowledge and PCK	Driving_question_generate_model	Conducts driving questions to encourage students to generate models (e.g., explanations) based on their prior knowledge.
Assessment interpretation and communication	Driving_question_curricular_model	Formulates a driving question and complements students' answers with a more sophisticated explanation or conceptual/curricular model.
Peer and self-assessment	E_model_utility_limitation_scope	Encourages students to evaluate their own models to help them identify the utility, scope, and limitations of the model they developed

It is worth mentioning that the R-ASMM was revised after having coded science teachers' assessment practices based on the transcription of classroom observations and interviews. In

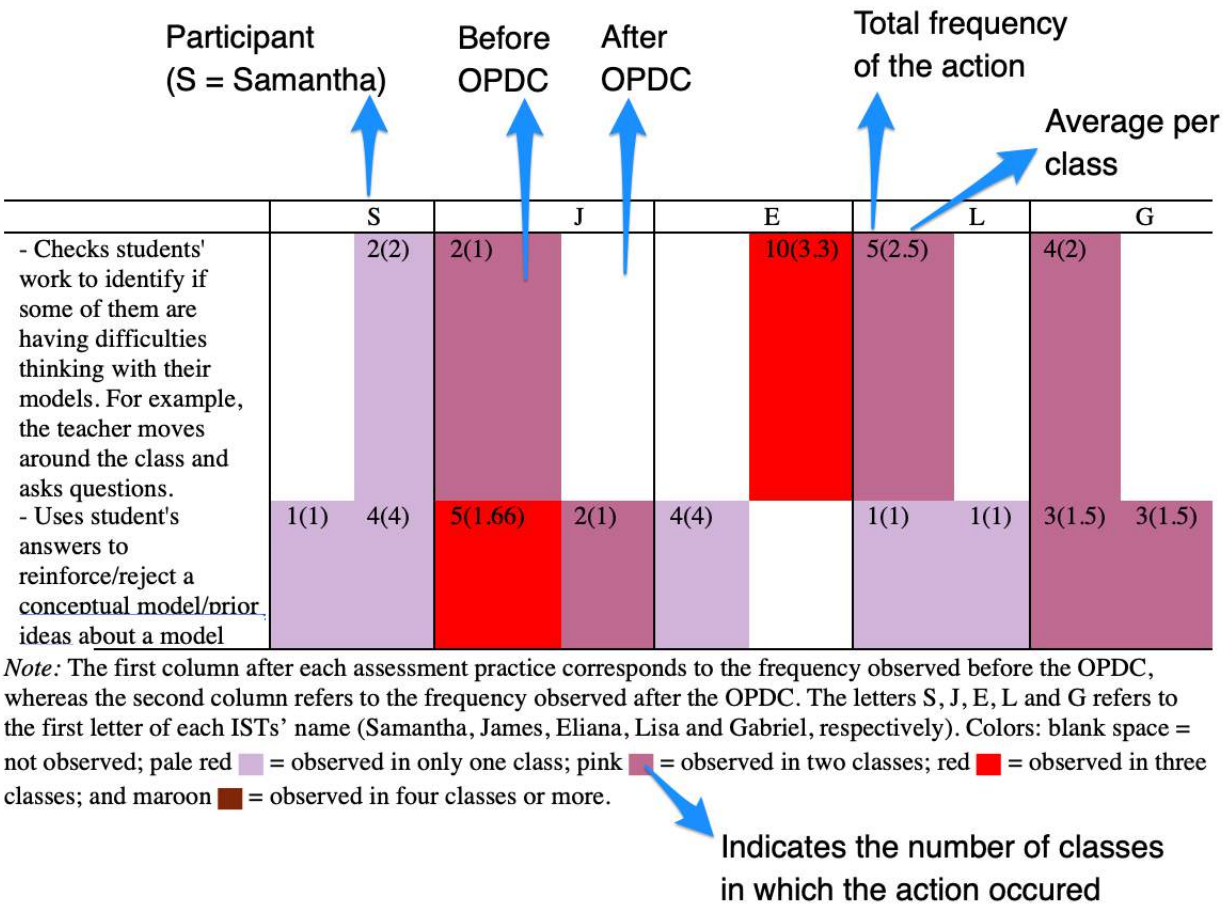
other words, the final version of the R-ASMM was administered once all the data from class observations for each IST had been coded. This decision was made to ensure that the rubric reflected science teachers' pedagogy instead of forcing the rubric to fit the data. In other words, the R-ASMM summarized the total frequency of ISTs' assessment practices for each of the dimensions included in ALMBT.

The R-ASMM was used to characterize teachers' assessment practices in their pedagogy and interviews. In the case of classroom observations and based on the analysis of the transcripts of the teacher-student interaction, for each dimension included in the theoretical framework in Figure 1 a table with the observed assessment practice was constructed to report the total number of instances (frequency) in which each type of assessment practice was observed during ISTs' pedagogy. To better visualize the frequency of each assessment practice, each table of the R-ASMM is detailed through a graphic heat map representing how many times that particular assessment strategy was observed across four-time points. Because in the case of Eliana, Gabriel and Lisa the number of observed classes was lower than four, whereas in the case of Samantha and James was larger than 4, I also report the average per class (in parenthesis) based on the number of class observations. The heat map indicates an ISTs' frequency of assessment practices over the unit observed in red scale from white through dark red (blank space = not observed; pale red = observed in only one class; pink = observed in two classes; red = observed in three classes; and maroon = observed in four classes or more). The R-ASMM included a table in which the first row showed the letter which represented the name of the participant "S" = Samantha, "J" = James, "E" = Eliana, "L" = Lisa, and "G" = Gabriel. Below each participant,

two columns were included which reflected the frequency in which each practice was observed in the class observation or mentioned in the interviews before and after the OPDC.

Figure 8

Example of R-ASMM for the Theoretical Dimension of Knowledge of Ethics of Assessment



Note: The first column after each assessment practice corresponds to the frequency observed before the OPDC, whereas the second column refers to the frequency observed after the OPDC. The letters S, J, E, L and G refers to the first letter of each ISTs' name (Samantha, James, Eliana, Lisa and Gabriel, respectively). Colors: blank space = not observed; pale red = observed in only one class; pink = observed in two classes; red = observed in three classes; and maroon = observed in four classes or more.

Figure 8 shows an example of R-ASMM for the theoretical dimension of knowledge of ethics of assessment. It can be noted that each column below each initial of the ISTs reflects the frequency of the strategy before and after the OPDC. For example, for the case of Samantha, only one strategy was identified before the OPDC ("Uses student's answers to reinforce/reject a conceptual model/prior ideas about a model"). This action occurred only one time and it was

observed only in one class before the OPDC (pale red) with a frequency of one time per class. In the second column, it can be noted that Samantha included both types of practices after the OPDC and it was also observed only in one class. The first number reflects the frequency of the action based on the analysis of the total number of classes observed after the OPDC, whereas the numbers in parenthesis reflect the average of the practice per class. In Samantha's case, the total frequency and the average are the same because the action was observed only in one class.

3.1.6.5 In-Service Science Teachers' Artifacts

Artifacts corresponded to data sources that teachers used to inform the narrative of their classroom and the instruments used by them to teach and assess models in the science classroom. Specifically, science teachers' artifacts included i) lesson plans which corresponded to a daily guide for what s/he would teach in a class, week, or unit; ii) summative exams (assessment instruments used to grade students' understanding or performance), and iii) class handouts and presentations.

3.1.7 Research Methods

3.1.7.1 Statistical Methods

Two types of statistical methods were used to analyze the data and answer the first research question related to whether ISTs' knowledge of models and modeling was related to their assessment literacy in MBT. Based on ISTs' responses in the QALMBT, descriptive statistics were used to present quantitative information in relation to science teachers' assessment literacy in MBT. Measures of central tendency (mean) and dispersion (standard deviation) were reported

for each of the dimensions included in the QALMBT-Modeling and QALMBT-Epistemic. Moreover, inferential statistics were used to infer from the sample data what was the population's assessment literacy in MBT. Specifically, linear regression and exploratory factor analysis were used which are detailed in the data analysis section. Linear regression was computed to determine the relationship between ISTs' knowledge of models and modeling and their assessment literacy, whereas exploratory factor analysis was used as a starting point for examining the underlying factors of the QALMBT related to ISTs' assessment literacy in MBT.

3.1.7.2 Classroom Observation Methods

Classroom observations were made involving video recordings. Video recordings facilitate the collection and analysis of observational data since they help the researcher to "video naturally occurring events that often elude the naked eye when seen in person but can become cleared upon review" (Schwartz & Hartman, 2007, p. 335). In this study, with permission from the teacher and after informing students' parents and guardians about the study, I set up a static camera positioned in one of the corners of the classroom. I decided this location to observe the whole classroom in order to have the possibility to select specific events related to the theoretical framework in Figure 1 (Derry et al, 2010). Between two and six consecutive lessons were recorded and transcribed for each IST for the class observations before and after the OPDC. I attended and recorded each session. The number of class observations fluctuated according to teachers' availability and different factors that affected the natural setting (e.g., extracurricular activities, political context, and exams). Students' responses were also considered in order to analyze teacher-student interaction in the classroom and analyze how ISTs engaged students in the use and development of modeling practices. Specifically, ISTs i) were observed teaching

science before the OPDC, ii) participated in a 1-hour semi-structured interview before and after attending the OPDC to explore how they reflected on their pedagogy, iii) attended an on-line professional development course in MBT, iv) participated in class observations after the OPDC, and v) participated in another interview after the OPDC. It is worth mentioning that on average teachers were observed between 4 and 5 times before and after attending the OPDC; however, the number of class observation after attending the OPDC was considerably lower for three of the participants due to the political context which occurred in Chile from mid-October to the end of December 2020.

3.1.7.3 Semi-structured Interviews

Two semi-structured interviews were conducted before and after the OPDC to further elucidate ISTs' ALMBT and explore how assessment influences their pedagogical decisions when teaching science. The interview protocol was first piloted with a Chilean chemistry and biology teacher who had almost ten years of teaching experience. Feedback on the interview protocol was received and some questions were reworded for clarity, simplicity, and answerability (Castillo-Montoya, 2016). The interviews before and after the OPDC were conducted in ISTs schools to answer the research question 2 and ascertain in what ways the OPDC had an impact on their approach to teach science, if at all. Each of the questions included in the interview protocol followed the dimensions in Figure 1 as areas to be explored regarding science teachers' assessment literacy in MBT. The full interview protocol is provided in appendix E. Examples of several questions asked in the interviews for the theoretical dimension of disciplinary knowledge are shown in Table 10.

Table 10

Examples of Questions Asked in the Interview Before and After Attending the OPDC

Theoretical Dimension	Interview before the OPDC	Interview after the OPDC
Disciplinary Knowledge and PCK	Could you give an example of how you include models in your class? In your classes, do you encourage students to build their own models? Yes: How do you motivate your students to create their own models? When students build a model, what is your objective of the activity?	Compared to the first round of observations before attending the online course, how did your ideas about how to include models in your classes change? If you were able to do so, how did you assess students' understanding of basic ideas in science using models.

Note: The questions asked in the second interview were slightly different to identify how ISTs' ideas had been reshaped after attending the OPDC.

The first interview was administered after the last class observation before ISTs attended the OPDC, whereas the second interview was conducted at the end of the second phase of observations once ISTs have participated in the OPDC. The script for both sets of interviews covered the eight dimensions used to conceptualize assessment literacy in MBT.

3.2 Data Analysis

This section presents how ISTs' assessment literacy in MBT and the identification and development of assessment literacy in MBT phase was analyzed. An explanation of methods used for the data analysis of the QALMBT is provided. A description of how thematic analysis was conducted is also provided. Finally, the next section covers how the qualitative information obtained from class observations, interviews, and ISTs' artifacts was analyzed and used to determine ISTs' levels of proficiency in MBT.

3.2.1 Data Analysis of In-Service Science Teachers' Assessment Literacy in MBT

This section presents how the quantitative data from the QALMBT questionnaire was analyzed by conducting an exploratory factor analysis and linear regression using *R* as the software environment for statistical computing. Issues related to the analysis of data are also explained which include i) the identification of the number of factors; ii) determination of model fit; iii) measures related to assessment literacy; iv) use of ordinary least-squares regression and v) response rate and handling missing data. Finally, an explanation of the thematic analysis of ISTs' assessment practices and the implementation of the rubric of levels of proficiency in MBT (R-LAPL) is detailed.

3.2.1.1 Construct Validity: Exploratory Factor Analysis

Factor analysis is a common method used to assess any evidence of construct validity (Besnoy et al., 2016; Thompson & Daniel, 1996). Markus and Lin (2012) define construct validity as "[W]hether the scores of a test or instrument measure the distinct dimension (construct) they are intended to measure" (p. 230) in order to support the interpretation and use of test scores. Factor analysis is a method that allows a researcher to investigate the internal structure of item responses (Markus & Lin, 2012) and "reduce the overall number of observed variables into latent factors based on commonalities within the data" (Atkinson et al., 2011). In science education, factor analysis is commonly utilized to support construct validity, see for example, Vishnumolakala et al., (2016) who used the Student Assessment of Learning Gains (SALG) instrument to measure latent constructs of students' assessment of their learning gains in knowledge and skills in chemistry.

To collect evidence to validate the internal structure of the QALMBT, I conducted an exploratory factor analysis (EFA). It is worth mentioning that I conducted three separate EFAs. Firstly, one EFA was conducted for the QALMBT-Epistemic because this section of the questionnaire was not related to assessment literacy and focused on exploring teachers' epistemological knowledge of models. Secondly, two separate EFAs were conducted for each version of the QALMBT-Generic and Modeling. I made this distinction in the questionnaire because each item stem was included twice to identify ISTs' assessment literacy in general versus assessment literacy in MBT. Moreover, I assumed that assessment literacy of a general kind might be different from an MBT approach; therefore, they might reflect different constructs.

Because the QALMBT is a new instrument that was developed to measure a new construct, "assessment literacy in MBT", EFA was chosen over confirmatory factor analysis (CFA) since the former is suggested to explore a new construct whereas the latter is often used to confirm the structure obtained in an EFA (Besnoy et al., 2016). EFA was performed to identify underlying dimensions within the QALMBT. This analysis can be used with the purpose to i) identify a latent variable, ii) measure dimensions that are related to a specific construct, and iii) conduct data reduction of variables into a more manageable number of factors (Field, 2009). The assumptions related to Kaiser-Meyer-Olkin (KMO) test of sampling adequacy and Bartlett's test of sphericity were applied to determine whether or not the variables could be grouped into a smaller set of underlying factors. These two analyses (KMO and Barlett's test), respectively, were required to determine the adequacy of the data for factor analysis (Hadi, Abdullah & Sentosa, 2016) regarding the sample size and whether there is a strong relationship between indicators used to define assessment literacy in this study (e.g., knowledge of ethics, knowledge

of grading). KMO is used as an "indicator of common variance within a data set, which indicates that latent factors may be present and EFA may be performed" (Howard, 2016, p. 52). For this test, values between .80 and .90 are ideal for performing a factor analysis (Field, 2009), and .60 can be used as a criterion for good factorability (Tabachnick & Fidell, 2001). These criteria are based on a general "rule of thumb" in which KMO is "estimated using correlations and partial correlations to test whether the variables in a given sample are adequate to correlate" (Diwivedi, 2007, p. 120). Regarding the Bartlett's test of sphericity, this test is conducted with the purpose of confirming the relationship between the variables (Diwivedi, 2007) and "checks whether the observed correlation matrix is an identity matrix" (Howard, 2016, p. 52). A *p*-value smaller than .05 for this test indicates that the test is significant, and that factor analysis can be conducted (Brace, Snelgar & Kemp, 2012).

An exploratory factor analysis or EFA was used to identify the underlying factors related to assessment literacy included in each version of the QALMBT questionnaire (Generic, Modeling, and Epistemic) using the *psych* package in *R*. Data for each section was screened for multivariate assumptions (normality, linearity, homogeneity, and homoscedasticity). The assumptions were met with slight problems of heteroscedasticity. Outliers were not eliminated from the data sample; however, I compared the results with and without outliers in order to identify if the results followed a similar trend. In the case of the sample without the outliers, thirty-seven multivariate were identified using Mahalanobis distance ($\chi^2(35) = 66.62, p < .001$), and they were removed from further analysis from the QALMBT-Generic and QALMBT-Modeling. A total of 27 participants were identified as multivariate outliers based on a cut-off value ($\chi^2(20) = 45.31, p < .001$), and they were removed from the sample. A total sample size of $N = 349$ was

included to conduct the EFA for the QALMBT-Generic and the QALMBT-Modeling, and a total sample size of $N = 345$ was used for the QALMBT-Epistemic. The results of the EFA for the data sample with outlier for each version of the QALMBT (Generic/Modeling/Epistemic) are presented in Tables 16 to 18. Only the results from the full data sample are presented and discussed in this section and the results for the data without outliers are shown in Appendix F. The determination of significant correlations through measures of sample adequacy showed that there was enough significant correlation to run a factor analysis. The measures of factorability were significant for the Bartlett's test of sphericity and KMO were greater than .8 (Field, 2009) for each of the sections of the QALMBT; Generic (KMO test indicated sampling adequacy, MSA = .94; Bartlett's test indicated correlation adequacy, $\chi^2(595) = 5507.80$, $p < = .001$), Modeling (KMO = .96; $\chi^2(595) = 7509.99$, $p < = .001$) and Epistemic (KMO = .9; $\chi^2(190) = 1899.56$, $p < = .001$).

3.2.1.2 Identification of the Number of Factors

To determine the number of factors retained from the EFA, different criteria were used, that are detailed below. The purpose of determining these factors was to identify the items in the QALMBT questionnaire that had a similar pattern of responses as a way to identify the components that might inform how teachers implement their assessment practices into their pedagogy. The first criterion was based on a rule of thumb that suggests that *eigenvalues larger than 1* should be retained (Kaiser, 1960). Another criterion suggested by Jolliffe (1986) indicates that *eigenvalues above .70* are recommended to be retained. Eigenvalues are computed by summing the squared factor loadings (Kline, 2014), that are used during the factor extraction methods. "A factor loading for a variable is a measure of how much the variable contributes to

the factor; thus, high factor loading scores indicate that the dimensions of the factors are better accounted for by the variable" (Young & Pearce, 2013, pp. 80-81). In EFA, "factors with small eigenvalues represent little common variances" (Howard, 2016, p. 53). Nevertheless, it is suggested to complement this information with other criteria since just judging by these cut-off values might result in overestimating the number of factors extracted (Costello & Osborne, 2005). In conjunction with the eigenvalues, it is also suggested to use the *scree plot test* (Yong & Pearce, 2013). Cattell's scree plot is a graph that shows the eigenvalues on the y-axis and the number of factors on the x-axis (Cattell, 1978; Smith & Alonso, 2020. The scree test is a subjective judgment that the researcher uses to identify the number of factors based on the visual scree plot (VSP) (Costello & Osborne, 2005). Specifically, this test is conducted by inspecting and interpreting the scree plot in order to identify a break or a bend in the plot where the curve flattens out (Smith & Alonso, 2020). The results are reliable when the sample size is at least 200 (Yong & Pearce, 2013). This point where the break occurs indicates the number of factors. An alternative to the scree plot and Kaiser's rule is conducting a *parallel analysis* (Horn, 1965). This technique offers one of the most promising results when determining the correct number of factors (Fabrigar et al., 1999). Horn's parallel analysis compares the observed eigenvalues to eigenvalues from random data (John et al., 2014), and assumes that meaningful factors extracted from the observed data must be larger than eigenvalues from random normal variates (Kaufman & Dunlap, 2000). In my study, I extracted the number of factors in the parallel analysis by using Principal Axis Factoring (PAF) and Maximum Likelihood Estimation (MLE). PAF "is a least-squares estimation of the common factor model. PAF makes no assumption about the type of error and minimizes the unweighted sum of the squares (unweighted least squares, ULS, or ordinary least squares, OLS) of the residual matrix", while MLE finds factors that maximize the

likelihood of producing the correlation matrix and "is derived from the normal distribution theory and assumes that all error is sampling error" (De Winter & Dodou, 2012, p. 696). Finally, another criterion was also used to check the dimensionality of the QALMBT. This criterion included determining the ratio of the first-to-second eigenvalues (Hattie, 1985). It is suggested that a ratio greater than four is evidence of unidimensionality (Reeve et al., 2007). Based on each of these criteria, I determined the dimensionality of the QALMBT and compared models with different factors.

3.2.1.3 Determination of Model Fit

After identifying the number of factors, factor analysis can be used to determine "[T]he extent that each variable represents each emergent factor through loading values" (Howard, 2016, p. 55), allowing the researcher to identify which variables are representative of a factor (e.g., items in a questionnaire) and which variables could be removed (Hinkin, 1998). Different authors have provided factor loading cut-offs to decide if a variable is representative of a factor. A good cut-off is often a factor loading equal to .40 (Hinkin, 1998), whereas other authors suggest values of .30 (Costello & Osborne, 2005; Gorsuch, 1983; Tabachnick & Fidell, 2007). Based on these criteria, the factor loadings of the items in the QALMBT were reviewed by the exploratory factor analysis to drop items with loadings less than .3 on a single factor and cross-loading items (Costello & Osborne, 2005).

Different measures were used to identify model fit based on Hu and Blenter' guidelines (1999). These fit indices included:

- The Tucker-Lewis Index (TLI), which "measures a relative reduction in misfit per degree of freedom" (Shi et al., 2019, p. 312).

- The Comparative Fit Index (CFI) which measures the "relative improvement in the fit of the researchers' model over that of a baseline model" (Kline, 2011, p. 208).
- The root mean square error of approximation (RMSEA), which is a "scale as a badness-of fit index where a value of zero indicates the best fit" with the lower and upper bounds of the 90% confidence intervals (Kline, 2011, p. 205).
- The root mean square of residuals (RMSR) which measures the average residuals for the correlation matrix.

Table 11 shows the values for model fit suggested by Hu and Bentler (1999) and Marsch et al. (2004). Each of these measures were used as indicators of improvements to explore new models that better fit the data (Hinton & Platt, 2019). In other words, the different models from the EFA were compared based on their goodness-of-fit indexes and analyzed and interpreted based on the theoretical framework included in Figure 1.

Table 11

Summary of Values for Model Fit Used to Compare the EFA of the QALMBT

Measure	Values for model fit		
	Poor	Acceptable	Good
Comparative Fit Index (CFI)	< .90	> .9	> .95
Tucker-Lewis Index (TLI)	< .90	> .9	> .95
Root Mean Squared Error of Approximation (RMSEA)	> .1	.06 -.08	< .05
Root Mean Squared Residuals (RMSR)	> .1	.06 -.08	< .05

The reliability of the factors generated from the EFA was evaluated through Cronbach's alpha. In science education, an arbitrary value of .70, or over .61 (Taber, 2018b) is often suggested as a sufficient index of the reliability of an instrument (Nunnally & Bernstein, 1994). Values closer to one and over .9 show excellent reliability. Finally, an oblique rotation method (oblimin) was used to conduct the EFA. This rotation method was chosen since it is commonly included in

social sciences research (Costello & Osborne, 2005) and has a history in science education research (Afari, 2015; Kind, 2007). Finally, the findings of the exploratory factor analysis are reported in the results.

3.2.1.4 Measures Related to Assessment Literacy

As was mentioned earlier, in order to consider a baseline of IST assessment literacy, I compared science teachers' responses based on a general approach to teaching science and an MBT approach. The items on the questionnaire include the same question twice (from a general approach when teaching science and from an MBT approach). I decided to call the scores obtained from the general approach as QALMBT-Generic, whereas the scores obtained from the items related to MBT were QALMBT-Modeling, both corresponded to dependent variables. The scores from the last part of the questionnaire related to ISTs' knowledge of the nature and purpose of models were called QALMBT-Epistemic and corresponded to the independent variable. It is worth mentioning that other independent variables were also included such as years of teaching experience and number of courses taken in assessment. To answer the first research question related to the relationship between ISTs' knowledge of models and modeling and assessment literacy in MBT, the variables to run the regression analysis are detailed in Table 12. As a result of the literature review and empirical correlation analysis on QALMBT, a set of variables was selected to study their relationship with ISTs' reported assessment strategies implemented in the science classroom. These variables corresponded to ISTs information provided in the first part of the questionnaire. whereas Table 12 provides information about the variables included in the regression analysis.

Table 12

Description of Variables Included in the Analysis

Variables	*Variable Name*	*Range*	*Measure Level*	Description
Outcome variables (Dependent variable)	QALMBT-Generic QALMBT-Modeling	0-175 0-175	Numerical	ISTs' QALMBT score
Assessment course	assessment_course	0-3	Numerical	Number of courses taken on assessment while studying the teacher education program
Topic Science Courses	topic_science_course	0-9	Numerical	A variable derived from a list of topics in science courses learned in their teacher education program or in professional development courses.
Years of teaching science*	year_experience	0-50	Numerical	A variable derived from ISTs' years of experience teaching science at middle/secondary level.
ISTs' knowledge about the nature and purpose of models*	QALMBT-Epistemic	0-100	Numerical	A variable derived from ISTs answer from the third section of the questionnaire related to their epistemological knowledge about models.

Note: * These variables were grand mean centered to help with the interpretation of the variables associated with the intercept (Hoffman & Gavin, 1998).

Multiple regression analysis is commonly used in education. "Once predictors in a regression model are selected, it is a common practice for researchers to investigate which predictors explains more variance than others, or to identify a sub-set of predictors that explain most of the variation in the outcome variable" (Liu et al., 2014, p.2). The following sections present how the linear models were obtained to investigate statistically if ISTs' understanding of models and modeling was related to ISTs' assessment literacy in MBT.

3.2.1.5 Ordinary Least-Squares Regression: Analysis of the QALMBT Questionnaire

Ordinary least-squares (OLS) regression was used as a technique to "[M]odel a single response variable which has been recorded on at least an interval scale. The technique may be applied to single or multiple explanatory variables and also categorical explanatory variables that have been

appropriately coded" (Hutcheson, 2011, p. 225). In order to investigate if ISTs' understanding of models and modeling was related to IST's assessment literacy in MBT, I performed an OLS regression. The multiple linear regression models for each version of the QALMBT questionnaire (QALMBT-Generic and QALMBT-Modeling) can be written as an equation 1:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_{2\ldots} \beta_n X_n + \varepsilon; \qquad n = 1, 2, 3, \ldots \tag{1}$$

where Y is the dependent variable (QALMBT-Generic or QALMBT-Modeling), X_1, X_2, *and* X_n are explanatory variables or independent variables used as predictors (e.g., QALMBT-epistemic, year_exp_science), n corresponds to the number of predictors, β_0 is the y-intercept (when x-axis = 0), β_n is the partial regression coefficient that represents the change in the mean value of Y when X increases by 1 unit when holding everything else constant (Abdi, 2003), and ε is the random error component.

For model selection, I used sequential regression (hierarchical or block-wise) since I did not rely only upon statistical results for selecting predictors. "Sequential entry allows the researcher greater control of the regression process. Items are entered and given order based on theory, logic, or practicality, and are appropriate when the researcher has an idea as to which predictors may impact the dependent variable" (Strickland, 2017, p. 88). The initial model only included the QALMBT-Epistemic as a predictor, and then, one predictor at a time was added to the model based on how these variables might impact teaching from an educational perspective (theoretically and logically), and therefore, impact the dependent variable. Each predictor was added sequentially based on the level of relevance that might impact the outcome. For example,

the variable topic_science_course was added to each model as the second predictor since those teachers who have learned more topics in science in their teacher education courses or professional development courses (e.g., assessing scientific reasoning, strategies to elicit students' ideas) might have a better understanding of the foundations of MBT and the inquiry practices in science. Then, the predictor assessment_courses was included in the model because I expected that ISTs' who took more courses on assessment in their teaching education programs might have a better understanding about the role of assessment in the science classroom. Years of experience was also included in the model because many prospective ISTs show an unsophisticated knowledge about scientific models and modeling even after finishing their four- or five-year degree programs (Danusso et al., 2010). Furthermore, more experienced science teachers often show a more diverse understanding about models but are still limited in this knowledge base (van Driel & Verloop, 1999). Finally, the predictor n_stud was added to the model because of the belief about a related strategy, inquiry is the time required to do it, and teachers who teach in large class sizes often struggle to find time to do inquiry (Llewellyn, 2013).

A common measure to identify if a model obtained from a linear regression is a good predictor of the dependent variable is the coefficient of determination R^2. This measure is the squared correlation coefficient in which values closer to 1 represent a better model, and values close to 0 represent no linear fit. "This statistic can be defined as one of the following: (a) the percentage of the variation that can be explained by the regression equation; (b) the explained variation divided by the total variation; or (c) the squared correlation coefficient (r)" (Harel, 2009, p. 1111). R^2 always increases when more predictors are added to a model (Harel, 2009). When adding new

predictors to a nested model, a common criterion used in sequential regression is the analysis of the adjusted R^2, which adjusts the R^2 based on the number of predictors and only increases if R^2 has some predictive value (Faraway, 2014). The comparison of adjusted R^2 was used as a first criterion to compare the generated models on the linear regression. To compare the fits of two models, analysis of variance (ANOVA) was conducted to analyze whether a more complex nested model was significantly better at fitting the data than a simpler model (Faraway, 2014). A significant *p*-value < .005 allows us to conclude that a more complex model significantly better fits the data than a simpler model. When the *p*-value is larger than .005, the simpler model must be preferred since it provides the best, parsimonious fit of the data.

Afterward, I performed a regression diagnostic. I checked assumptions of linear regression (Casson & Farmer, 2014; Mertler & Reinhart, 2016) such as the independence of errors through the Durbin-Watson test, normality (residuals are normally distributed) by looking at histograms and Q-Q plot, linearity by exploring the scatterplots of standardized residuals against fitted values and heteroscedasticity through the studentized Breusch-Pagan test (Breusch & Pagan, 1979). The model diagnostics of the final models showed that almost all the assumptions were met for the Canadian and the Chilean sample. The values of the residuals were independent with values in the range between 1.5 and 2.5 for the Durbin-Watson test; the residuals were slightly negatively skewed for each model with the exception of the QALMBT-Generic for the Canadian sample, which was normally distributed; the Q-Q plot included some data points that hardly touch the line which might suggest that the normality of the residuals may have been violated. It is worth mentioning that the assumption of normality can be relaxed in a large sample size since the non-normality of residuals might not adversely affect the inferential procedures because of

the central limit theorem which assumes that the sampling distribution of the sample means tend to a normal distribution (Pek et al., 2018). The scatterplots of the standardized residuals against fitted values showed that the residuals were mostly equally distributed meaning that the homogeneity of variance was held, and finally, the Breusch Pagan Test was not significant; therefore, the null hypotheses that the variance of the residuals is constant can be accepted suggesting that the data was homoscedastic. It is worth noting that for the visualization of the multiple regression model, I only analyzed scatter plots. Because multiple predictors were included in the regression models, I did not explore 3-D plot representations that can be used when one dependent variable and two explanatory variables are included in the regression model. An alternative option to visualize the predictors included in the regression model involves the use of conditional plots, also called coplot, a graphical method in which a subset plot includes two variables conditional on the value of a third variable. Nevertheless, the interpretation of the equation of the regression models is simpler to visualize and interpret in comparison to the analysis of a subset of plots with multiple predictors.

Cooks' distance was used as a general rule of thumb to identify influential observations. Influential observations correspond to those observations that might cause a substantial change in the estimates of the coefficient model if they are removed (Zhang, 2016), and Cooks' distance is a measure of influence with a standard 0.5 threshold. Influential points based on Cook's distance were not identified in the data. Finally, the degree of collinearity of the predictors included in the model was assessed by analyzing the Variance Inflation Index (VIF) statistic. "VIF indicates the strength of the linear dependencies and how much the variances of each regression coefficients is inflated due to collinearity compared to when the independent variables are not linearly related"

(Yoo et al., 2014, p. 10). VIFs greater than 2.5 usually suggest a problem of collinearity, which indicates that an independent variable is highly correlated to another independent variable (Allison, 1999). There was no multicolliniarity problem in the data since the VIF values were close to 1. The results of the OLS for the QALMBT-Generic, Modeling, and Epistemic are revealed in the results chapter to answer the first research question of this study.

3.2.1.6 Response Rate and Handling of Missing Data on QALMBT

Of the total of 45 ISTs in Canada who answered the questionnaire, 43 of them finished the whole questionnaire. In the case of Chile, 398 ISTs completed 34% of the QALMBT-Generic and QALMBT-Modeling. From these 398 ISTS, 386 ISTs completed two-thirds of the questionnaire. In other words, participants answered the whole section related to QALMBT-Generic and QALMBT-Modeling. Finally, from the initial sample of 398 ISTs, 372 ISTs completed the full version of the questionnaire. Listwise deletion was used to delete ISTs' answers that were not completed. This strategy involved the removal of cases with missing values (Enders, 2006). The removal of 5% of the sample is acceptable for statistical inferences (Schafer, 1999). Hence, two samples were considered for analysis. The first sample included those ISTs who completed each question of the QALMBT-Generic and the QALMBT-Modeling. The second sample included those ISTs who completed each section of the questionnaire.

3.2.1.7 Thematic Analysis of ISTs' Assessment Practices

The use of thematic analysis in science education is common (e. g., Beck et al., 2020; Patron et al., 2017;) when analyzing qualitative data, which in my study corresponded to the data collected from classroom observation, interviews and ISTs' artifacts. Nowell et al. (2017) suggest six steps

when using thematic analysis. These steps include; i) familiarizing with the data, ii) generating initial codes, iii) searching for themes, iv) reviewing themes, v) defining and naming themes, and vi) producing the report. Similarly, Yin (2011) and Castleberry and Nolen (2018) emphasize that thematic analysis clearly includes five stages which are i) compiling, ii) disassembling, iii) reassembling, iv) interpreting and v) concluding. These five stages suggested by Yin (2011) were used in my study to conduct a thematic analysis on my data. Triangulation was pursued by analyzing these different data sources collectively to support inferences about what teachers know about models and modeling and how they assess students while guiding them in the process of generating, using, testing, and modifying models.

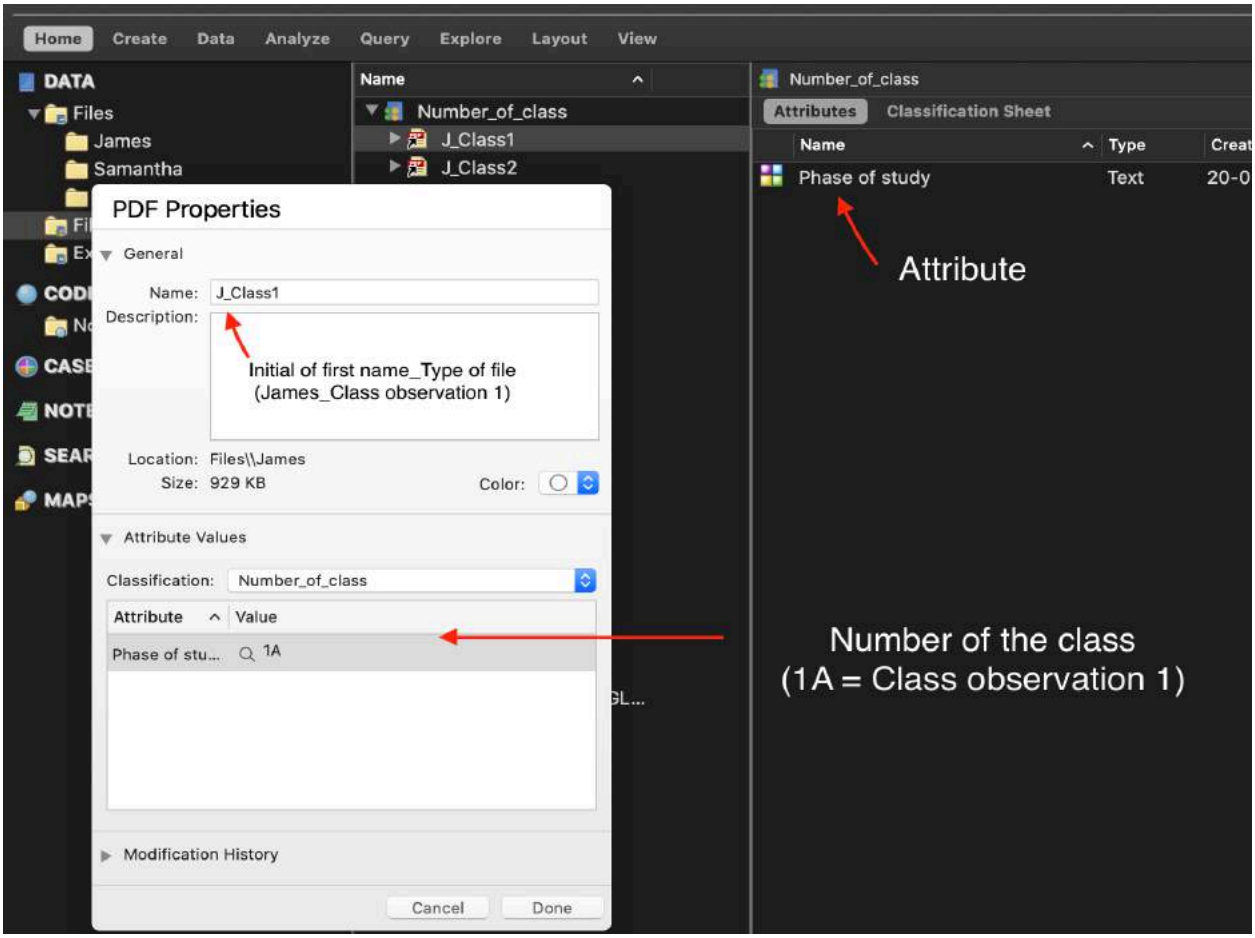
The first step in thematic analysis, *compiling*, is related to organizing the data into a formal database (e.g., transcribe the data) and the familiarization with the data by reading the transcripts several times (Vaismoradi et al., 2013). In my study, I took an active role during the process of data collection and data analysis. I collected, transcribed, and analyzed the data, which helped me get a sense of the quality of the data and how this might be useful to answer the research questions. The transcription process spanned over three months, and each class observation took an average of 6-8 hours to be transcribed, and the transcriptions of each interviews took 4-5 hours. I typed up the transcriptions using Microsoft Word, and the audio and video files were transcribed verbatim. The transcribed data was read and re-read, which gave a better sense of the information. Seven hundred and fifty-one pages of textual data were generated from the class observations and interviews.

The data was imported into in NVivo 12 for the thematic analysis. Based on the observational rubric (R-ASMM), ISTs' assessment practices from the analysis of the transcripts from interviews and class observations, and their artifacts (exams and lessons plans) were thematically analyzed. Each interview was transcribed. and coded using a codebook (Examples of codes included in the codebook are provided in Figure 13). Similarly, class observations were also transcribed and coded. The answers were transcribed into Microsoft Word and then analyzed via NVivo. Ellipses, which corresponded to parenthesis with three dots (…), were used to express an omission in the text when presenting excerpts from interviews of student-teacher dialogues. Three dots without the parenthesis were used to indicate that the teacher was thinking an answer or thought for a couple of seconds before continuing with an explanation. It is worth mentioning that I translated from Spanish into English each of the excerpts included in this dissertation. The translation of two full samples of data (one interview and one class observation) were double-checked by two independent reviewers who were fluent in Spanish and English. Examples of the codes included in the codebook and used by the reviewers are detailed in Figure 13. A constant comparative method was used to identify patterns from the data from class observations and interviews which were classified by using the option from NVivo called "attributes" (see Figure 9). Attributes allow the researcher to classify a specific piece of information, source, or file by tagging each source. I created an attribute called "Phase of study" in which I classified each file based on the moment in which the data was collected, for example, the transcriptions of the first class observations were tagged as "1A", the transcriptions of the first class observation after attending the online professional course were tagged as "1B", the first interview before the OPDC were tagged as "Pre_interview" and the last interview after attending the online professional development course was tagged as "Post_Interview" (see Figure 11). The attribute

generated to organize the data was used later to compare the frequency and type of code identified in each source.

Figure 9

Example of Folders and Attributes Used in NVivo to Organize the Data

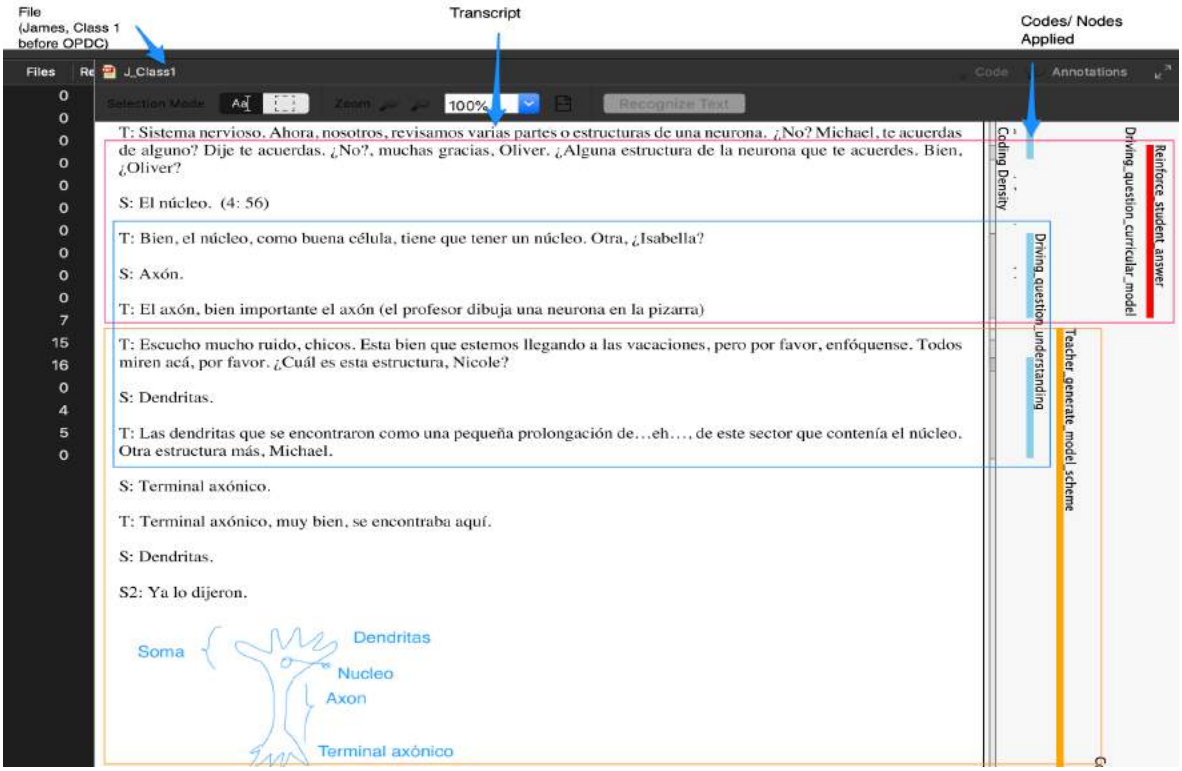


The second phase, *disassembling*, is related to breaking down the data into smaller fragments in order to generate meaningful groupings, for example, by the process of coding. Sutton and Austin (2015) define coding as "[T]he identification of topics, issues, similarities, and differences that are revealed through the participants' narratives and interpreted by the researcher (p. 228). A code in this study (generally referred as nodes in NVivo), refers to a short phrase or

concept that capture “the key aspect as important in that bit of data” (Clarke et al., 2015, p. 235). The data from the phase of identification and development of assessment literacy proficiency was coded in order to respond to research question 2 regarding in what ways ISTs’ assessment literacy about models and modeling influence their pedagogy and how their pedagogy changed as a result of attending an online professional development course on MBT. Figure 13 illustrates an example of relevant codes that were applied to an excerpt of the class observation. In my study, the process of coding involved inductive reasoning (Charmaz, 2006) and an iterative process during the constant comparative analysis (Chun Tie et al., 2019), as is suggested by Strauss and Corbin (1998). MBT was analyzed following an iterative process of developing codes. For example, Figure 10 shows an example of coding.

Figure 10

Example of how Coding was Conducted to the Transcripts

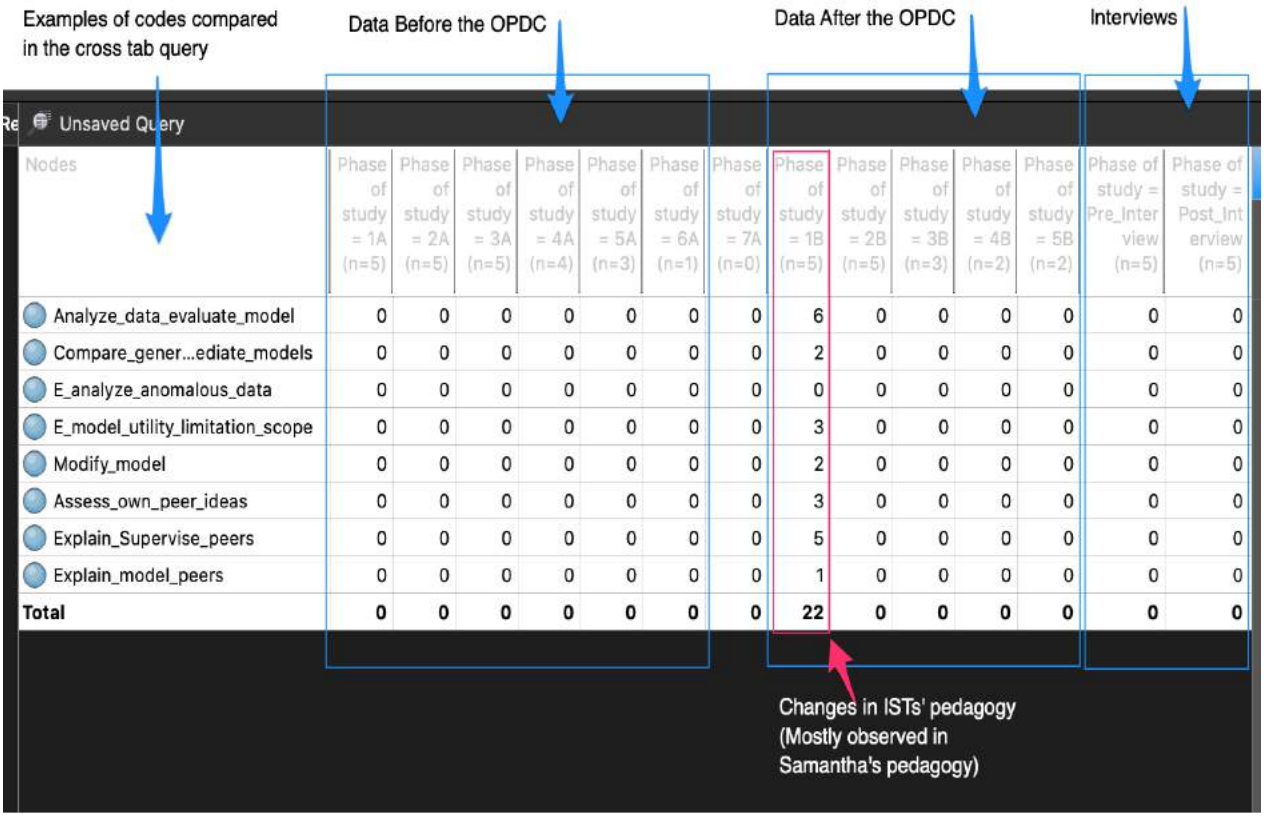


In Figure 10 three codes can be identified. Each color in the excerpt represents a code (assessment strategy implemented by the teacher). For example, the code in orange, "teacher generate_model_scheme", indicates that the teacher drew a scheme to represent the features of the target model (neuron structure model). The third phase of coding, *reassembling,* involves a procedure in which codes or categories are rearranged and put together to create potential themes (Scharp & Sanders, 2018); for example, through thematic hierarchies that help to visualize the clustering of similar codes and higher-order codes or by looking for patterns. During the iterative process of generating, reviewing, and refining codes, all of the qualitative data (interviews, class observations, and science teachers' artifacts) was explored until reaching saturation in terms of new themes. Double-blind coding occurred with a member of the supervision team and helped to foster greater trustworthiness in the coding. Six cycles of revisions of the codes were conducted until reaching agreement through discussion sessions. After coding the data, I searched for patterns in a cross-case analysis (comparison among ISTs' assessment practices in their pedagogy) and clusters of key concepts, and subsequent sub-concepts were elaborated based on the theoretical dimensions of ALMBT used in this study (see Figure 1). In this process, the data from each participant was compared among the others. I used crosstab queries in NVivo as an efficient way to represent the spread of coding by participants. For example, Figure 11 shows that major changes in ISTs' pedagogy occurred in the first class after the OPDC in Samantha's case. The column in red shows the type of assessment strategies implemented by the teacher (e. g., analyze_data_evaluate_model, E_analyze_anomalous_data) and the frequency in which each strategy was observed in ISTs' pedagogy. The number 22 (last row) represents the total number of instances in which the assessment practices for the whole dimension were observed. This

figure shows an aggregated crosstab query for the dimension of "knowledge of peer and self-assessment" for Samanatha's case.

Figure 11

Example of Crosstab Query for the Dimension of Knowledge Peer and Self-Assessment



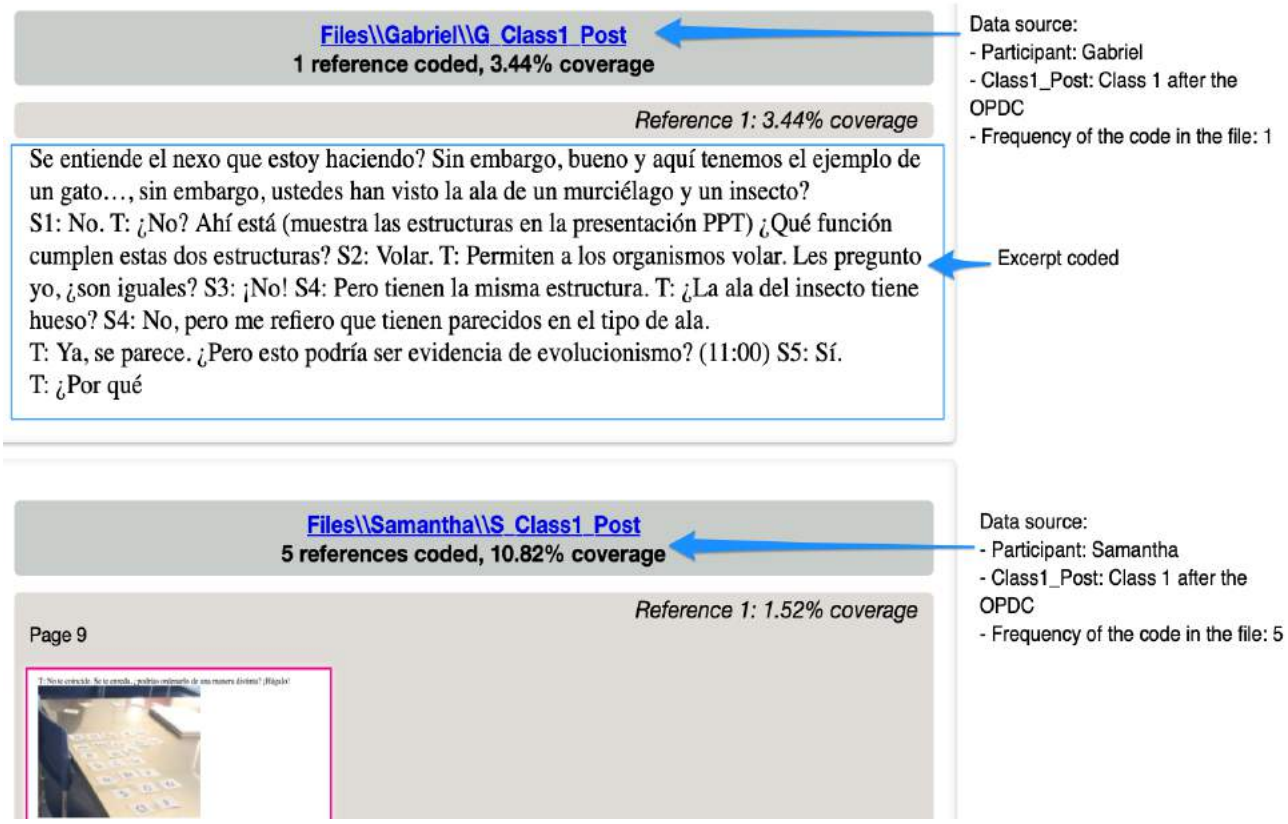
Nodes	Phase of study = 1A (n=5)	Phase of study = 2A (n=5)	Phase of study = 3A (n=5)	Phase of study = 4A (n=4)	Phase of study = 5A (n=3)	Phase of study = 6A (n=1)	Phase of study = 7A (n=0)	Phase of study = 1B (n=5)	Phase of study = 2B (n=5)	Phase of study = 3B (n=3)	Phase of study = 4B (n=2)	Phase of study = 5B (n=2)	Phase of study = Pre_Interview (n=5)	Phase of study = Post_Interview (n=5)
Analyze_data_evaluate_model	0	0	0	0	0	0	0	6	0	0	0	0	0	0
Compare_gener...ediate_models	0	0	0	0	0	0	0	2	0	0	0	0	0	0
E_analyze_anomalous_data	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E_model_utility_limitation_scope	0	0	0	0	0	0	0	3	0	0	0	0	0	0
Modify_model	0	0	0	0	0	0	0	2	0	0	0	0	0	0
Assess_own_peer_ideas	0	0	0	0	0	0	0	3	0	0	0	0	0	0
Explain_Supervise_peers	0	0	0	0	0	0	0	5	0	0	0	0	0	0
Explain_model_peers	0	0	0	0	0	0	0	1	0	0	0	0	0	0
Total	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**22**	**0**	**0**	**0**	**0**	**0**	**0**

After coding teachers' assessment practice, for example, for the code "analyze_data_evaluate_model," I reviewed the frequency that each IST performed the specific action. In figure 11 it can be noted that this action appeared 6 times in total in Samantha's pedagogy. This situation occurred because codes related to teachers' instructional practices were coded each time that a student or a teacher formulated or answered a question. It is worth mentioning that I only counted the same action again if during student-teacher interaction the

event involved a new topic or a change in how the dynamic of the class occurred. For example, student 1 asked a question related to idea 1 for topic 1, and another student asked a different question about idea 2 for the same topic (coded twice). If student 1 and 2 continued discussing the same idea with the teacher, the whole event was coded once. The Crosstab Query, which in other words represented the R-ASMM, since the rubric includes the explanation of each code and the frequency that each assessment strategy occurred, was also used to inform the type of assessment strategies that teachers more often implemented before and after the OPDC. For example, Figure 11 shows that 8 different types of assessment strategies were implemented by Samantha only in the first class after the OPDC. Figure 12 shows an example of coding.

Figure 12

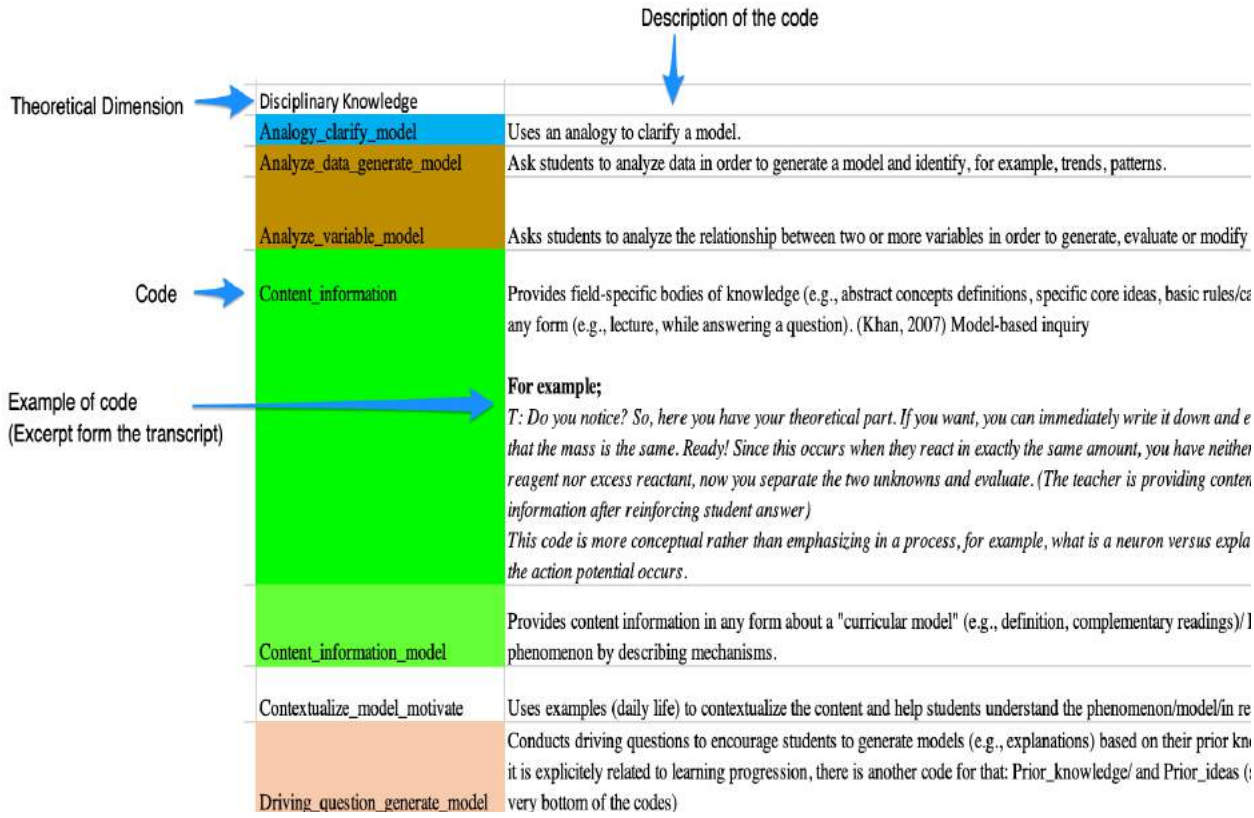
Example of Coding



Axial coding was used to identify similarities and differences among cases (ISTs). Through the analysis of the nodes generated in NVivo, representing the codes, initial themes and subthemes were generated, and the definition of each theme and node was constantly revised (see example of codebook in figure 13).

Figure 13

Example of Codebook for the Theoretical Dimension of Disciplinary Knowledge and PCK

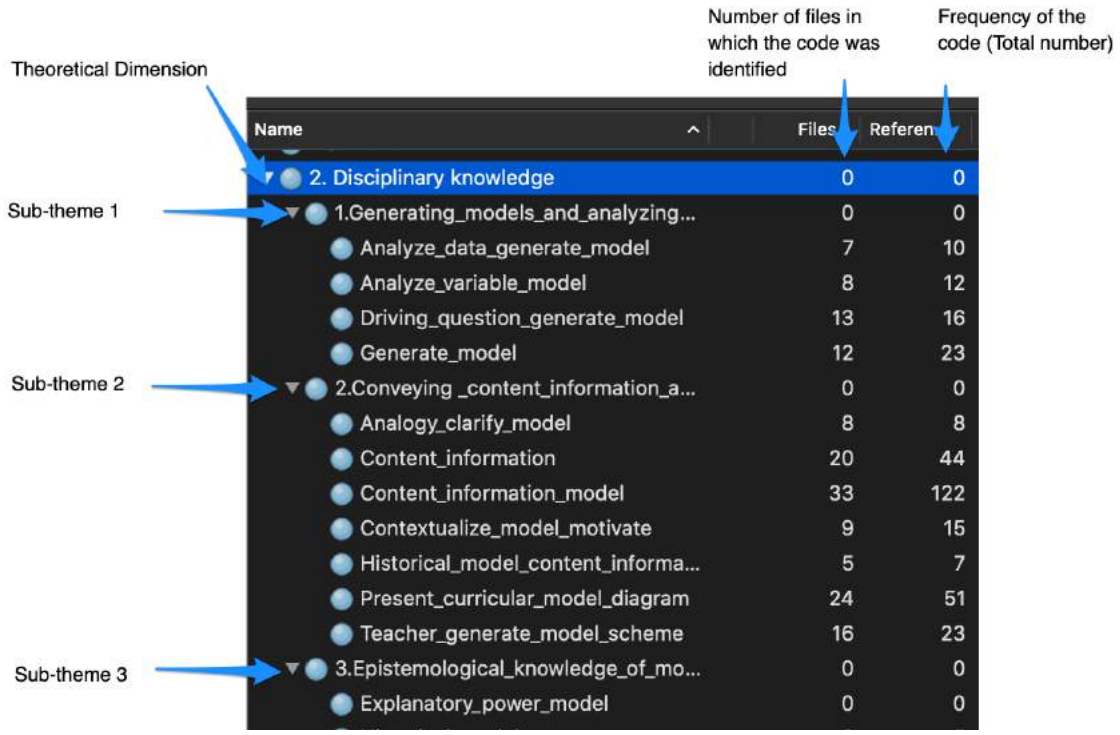


When including thematic hierarchies, Clarke et al. (2015) suggest that hierarchies should not include more than three theme levels. These three hierarchical levels include i) *overarching themes* (a theme that organizes and captures a major number of themes) (Braun & Clarke, 2013), ii) *themes* (it is a central organizing concept or idea that is analyzed from the data), and iii) *sub-themes* (it is a specific idea that is captured and developed from a theme). Figure 14 shows an

example of sub-themes that emerged for the theoretical dimension related to "disciplinary knowledge and PCK".

Figure 14

Example of Axial Coding and Generation of Themes



The fourth phase of *interpretation* involves the process of finding how themes are related to each other and establishing a relationship between them. It is also worth noticing that thematic analysis has been criticized in qualitative research for lacking of "[A] pragmatic process for conducting trustworthy thematic analysis" (Nowell et al., 2017, p. 2). In order to conduct a rigorous analysis, I followed the 15-point checklist suggested by Braun and Clarke (2006) as criteria for good thematic analysis. Figure 15 shows the steps followed to conduct the thematic

analysis and Table 13 indicates the criteria suggested by the author, and how they were achieved in this study in the right-hand side column.

Figure 15

Steps Followed to Conduct the Thematic Analysis

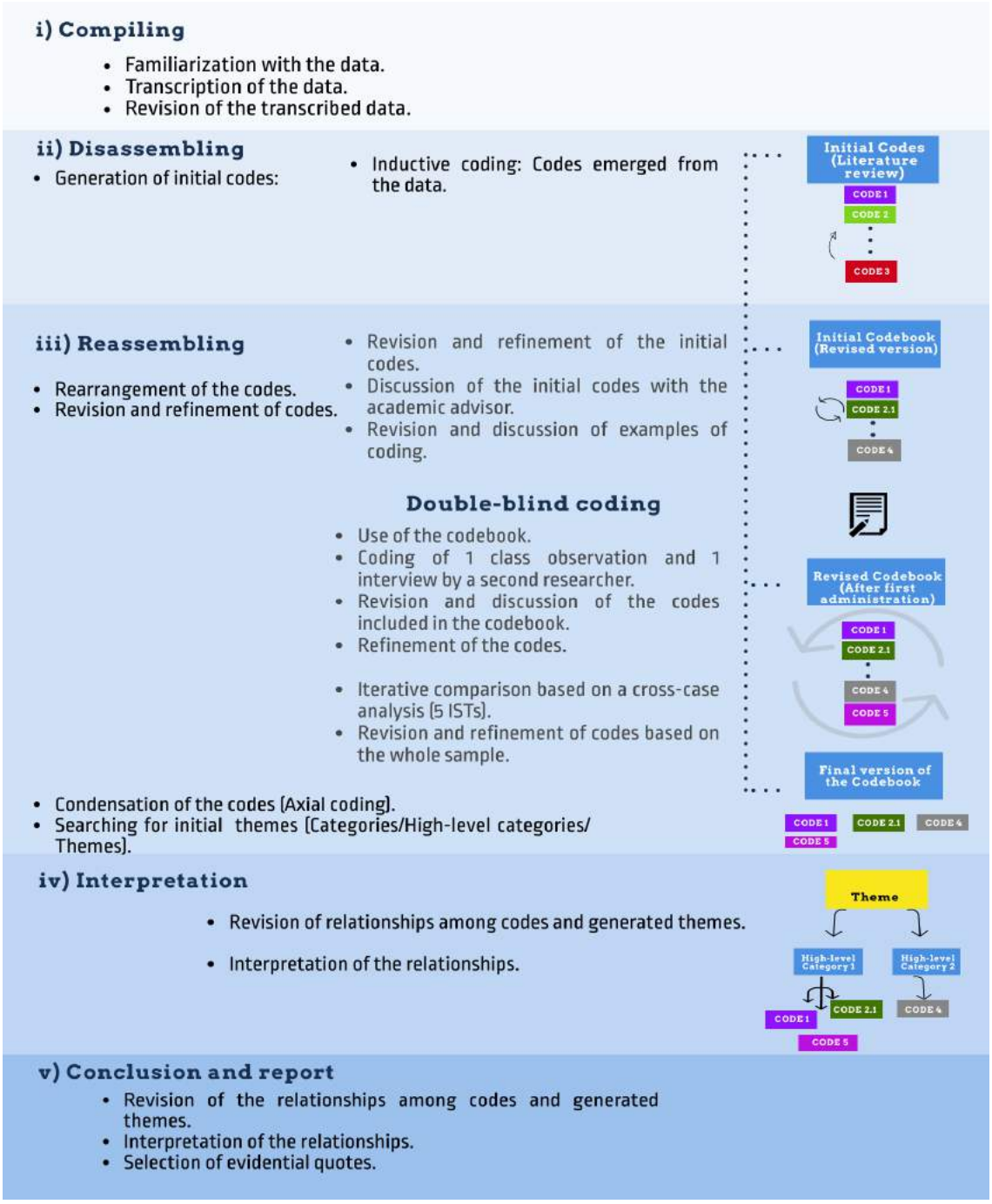


Table 13

Criteria for Good Thematic Analysis Used in this Study

Process	Criteria	How the criteria were achieved in this study
Transcription	1. The data have been transcribed to an appropriate level of detail, and the transcripts have been checked against the tapes for 'accuracy'.	1. I transcribed all the data and conducted several checks during the process of revision of the audio files.
Coding	2. Each data item has been given equal attention to the coding process.	2. The entire transcript for each participant was coded.
	3. Themes have not been generated from a few vivid examples (an anecdotal approach).	3. The themes were generated from the analysis of different pieces of information. Examples are provided to support each claim.
	4. All relevant extracts for each theme have been collated.	4. Extracts were aggregated and included in the nodes created in NVivo.
	5. Themes have been checked against each other and back to the original data set.	5. The themes were revised and compared against each other for agreement within all the data.
	6. Themes are internally coherent, consistent, and distinctive.	6. Themes were revised and guided by the research question of this study.
Analysis	7. Data have been analyzed - interpreted, made sense of - rather than just paraphrased or described.	7. Themes were analyzed based on the existing literature related to the purpose of this study.
	8. Analysis and data match each other – the extracts illustrate the analytic claims.	8. Through an iterative process of revision, the themes and codes were analyzed and compared. Clear examples from transcriptions are indicated in the findings.
	9. Analysis tells a convincing and well-organized story about the data and topic.	9. The themes are based on ISTs' observation and their ideas and provide evidence of participants' assessment literacy in MBT.
	10. A good balance between analytic narrative and illustrative extracts is provided.	10. The themes are supported by evidential quotes that were used to interpret the data.
Overall	11. Enough time has been allocated to complete all phases of the analysis adequately without rushing a phase or giving it a once-over-lightly.	11. An initial idea of the data was obtained during the process of class observation and transcription. Moreover, the analysis of each piece of information was conducted separately and on different days.
Written report	12. The assumptions about, and specific approach to, thematic analysis are clearly explicated.	12. How thematic analysis was conducted to analyze the data was clearly informed throughout this study.
	13. There is a good fit between what you claim you do, and what you show you have done – ie, described method and reported analysis are consistent.	13. The methods and research design were clearly informed and followed during the collection and analysis of the data.
	14. The language and concepts used in the report are consistent with the epistemological position of the analysis.	14. Interpretivism was used as an approach to explore and analyze the data. ISTs' levels of proficiency in assessment literacy in MBT were interpreted based on participants' voice and pedagogy.
	15. The researcher is positioned as *active* in the research process; themes do not just 'emerge'.	15. My active role in the study was clearly informed, and the findings were continuously revised during the process of data analysis.

When conducting the analysis, I also covered some of the core elements or considerations suggested by Weed (2009) to evaluate qualitative research's internal consistency. Even though these core elements are suggested for grounded theory, these elements are still applicable, in Weed's view, for studies that implement other qualitative methods, such as thematic analysis. These core elements to evaluate for internal consistency include:

- An iterative process: In this study, the data was regularly compared with the literature in MBT and the literature in assessment literacy to review and refine the codes that were included in the codebook.
- Theoretical sampling: Glaser and Strauss (2006) define theoretical sampling as "[T]he process of data collection for generating theory whereby the analyst jointly collects, codes, and analyzes his data and decides what data to collect next and where to find them, in order to develop his theory as it emerges" (p. 45). Even though the sample selected in this study may seem based on structural circumstance, the criteria used were based on a theoretical purpose where cross-case analysis permitted a comparison of different sources of data such as teachers' artifacts, lesson plans, and interviews.
- Theoretical sensitivity: Theoretical sensitivity is subjected to the researcher's personal and professional experiences; therefore, I tried to be flexible to different ideas that were developed from the data. I was aware of the literature in the field, and I had an open mind to explore and question the data (Charmaz, 2006) in order to evaluate the inferences and assumptions that I made from participants' answers and experiences.
- Codes: inductive coding was used as the analytic process to explore the data in which the categories and sub-categories were developed from the data (Blair, 2015) and included in the codebook which was constantly revised.

- Constant comparison: Constant comparison is a method that allows us to make conclusions and inferences based on a systematic revision of codes and the comparison between incidents that are applicable to each category (Glaser & Strauss, 2006). In my study, the finding from each participant was compared among them iteratively to check whether the codes and concepts generated remained relevant.
- Theoretical saturation: Saturation is achieved when "[G]athering fresh data no longer sparks new theoretical insights, nor reveals new properties of these core theoretical categories" (Charmaz, 2006, p. 113). While collecting data from class observations after attending the online professional development course, I observed that teachers tended to follow similar patterns in their practice as they did during the first cycle of class observation before the course. I must acknowledge that new ideas emerged among participants' responses during the second interview, but many of these ideas tended to repeat among participants and did not reflect a new insight that might have required to continue with further interviews. Through an iterative process of coding and re-coding data, theoretical saturation was achieved after analyzing the entire data set and no new categories appeared to emerge.

3.2.1.8 Rubric of Levels of Proficiency in Assessment Literacy (R-LPAL)

In order to characterize ISTs' proficiency in assessment, I developed an R-LPAL rubric (Appendix G) which was used as a form of analysis to explore how ISTs' assessment literacy changed after the OPDC. I adapted the Model-Based Inquiry learning progression of teaching performance proposed by Furtak et al. (2012) and incorporated several ideas from the dimensions suggested by Bennett (2017) about generative thinking, metacognition and causal

reasoning in MBT. In other words, I adapted and extended the Furtak et al. rubric to identify how teachers immerse and assess their students in the generation, evaluation, and modification of models through the engagement of modeling practices. The indicators included in the R-LPAL were organized based on Figure 1 and the theoretical dimensions indicated in the knowledge base of assessment literacy in practice. The R-LPAL rubric included 4 levels of proficiency for teachers: i) novice, ii) advance beginner, iii) competent, and iv) advanced assessor. These levels indicate how sophisticated is the ISTs' pedagogy to implement and assess students using an MBT approach. R-LPAL, was used to describe the complexity of ISTs' assessment literacy according to Figure 1 and compare ISTs' pedagogy before and after the OPDC. Table 14 shows an example of levels of teacher proficiency in terms of their disciplinary knowledge and PCK when teaching with models. Qualitative judgments were made for the overall observation, one before the OPDC and one after the OPDC.

Table 14

Examples of Teacher Indicators Included in the R-LPAL

Novice	Advanced beginner	Competent	Advanced assessor
Disciplinary Knowledge and PCK			
Follows a lectured-based approach in which models are used to complement a definition and/or define components/features of a system (e.g., students observe a representation of an animal cell and the teacher defines the structures).	Uses models in the class mostly to present concepts, ideas, theories (e.g., the teacher presents and explain a curricular model, for example, a heart and lung diagram).	Engages students in activities that involve the generation of a model; however, the generated model is barely used in the classroom.	Engages students in the generation of a model which is used as a research tool to generate information and understand a mechanism or phenomenon.

Note: Only one example for the dimension of disciplinary knowledge and PCK is shown in the table.

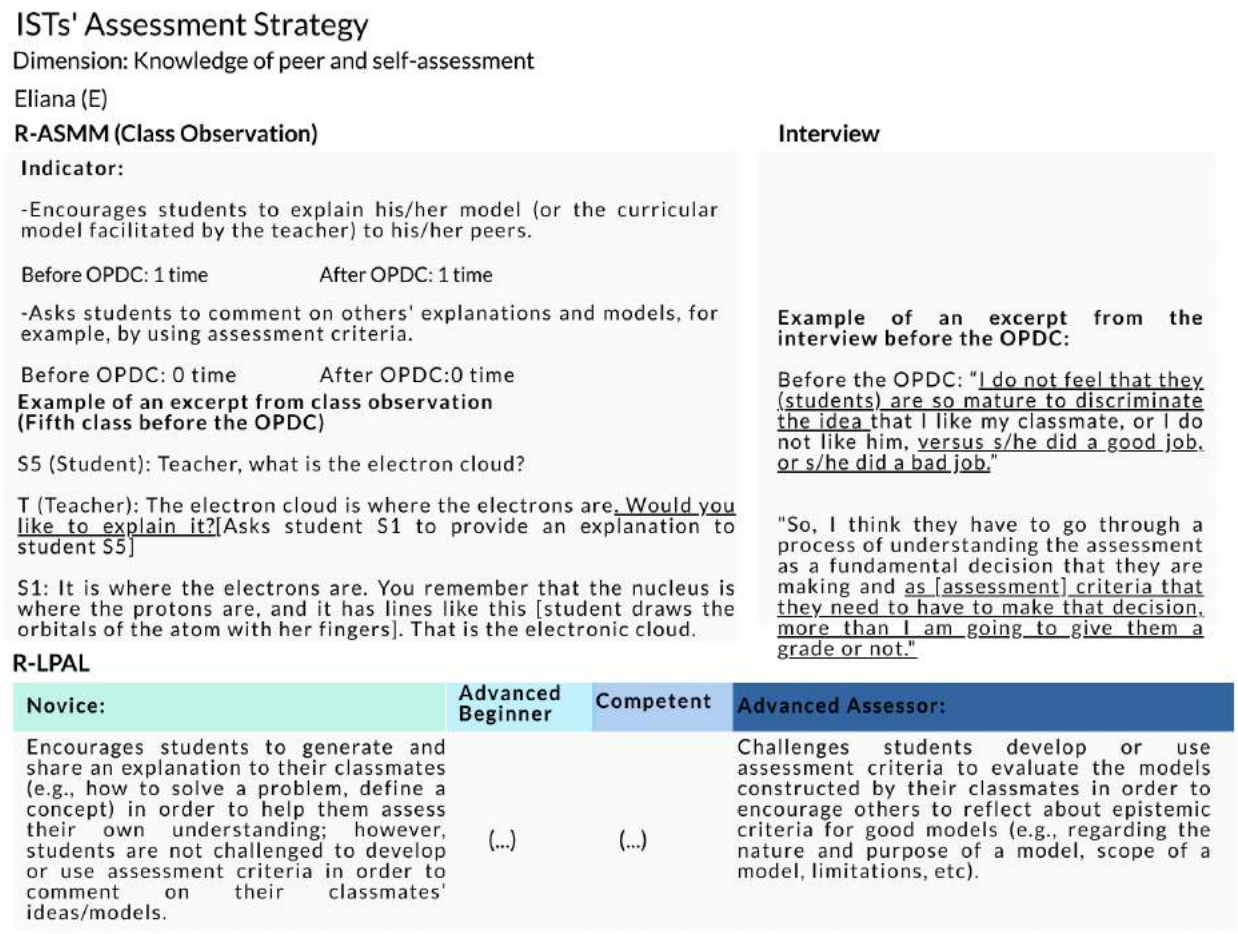
The R-LPAL was applied before and after the OPDC, and also to explore whether there was a progression in how ISTs' assessed student's modeling practices during the class observation. To better visualize how science teacher's assessment literacy about models and modeling influence

their pedagogy (in terms of their levels of proficiency) (research question 2), I also used heat maps when implementing the R-LPAL as a tool to facilitate the visualization of patterns (Yu and He, 2017) before and after the OPDC. Unlike the R-ASSM, the heat map used in the R-LPAL does not refer to the frequency in which the indicator was observed, instead it refers to ISTs' level of achievement of the indicator. A blue scale was used to characterize each level of proficiency aqua blue = novice; cyan blue = advanced beginner; blue = competent; dark blue = advanced assessor). The reason for the color change (blue scale) was for data visualization purposes to depict the levels and to avoid confusion with the R-ASMM that used red shading to represent the number of classes in which the action occurred. The use of the R-LPAL occurred after having analyzed the different data sources. For example, after the process of coding science teachers' assessment practices based on observable actions, the frequency, and type of assessment strategies were captured by the R-ASMM and then added with other data sources such as the transcriptions of interviews and ISTs' artifacts. During the methodological triangulation, difference data sources were analyzed to establish the level of proficiency in the R-LPAL. It is worth mentioning that a teacher could have been scored as an advanced assessor even if the action was observed in their pedagogy only once before or after the OPDC. Therefore, when using the R-LPAL I tried to keep the full range of levels of proficiency by dimension in mind as I judged ISTs' performances. Based on the data set, the levels of information related to the declarative level from interviews, class observations planning and design level (lesson plans and artifacts) were used to rate ISTs' pedagogy. The results from the QALMBT were used as a baseline to identify changes among ISTs after attending the OPDC. Using Eliana as an example, Figure 16 shows a diagram that summarizes the process of scoring

ISTs' level of proficiency for one of the indicators related to knowledge of peer and self-assessment.

Figure 16

Example of Methodological Triangulation to Label ISTs' Level of Proficiency



To rate science teachers' level of proficiency the data sources (QALMBT questionnaire, R-ASMM, transcriptions of interviews and IST's artifacts) were jointly analyzed and coded in the R-LPAL. For instance, for the theoretical dimension of knowledge of peer and self-assessment, the R-ASMM showed that none of the ISTs' challenged their students to develop or use assessment criteria to evaluate the models constructed by their classmates in order to encourage

others to reflect on epistemic criteria for good models. From Figure 16 (left side), it can be observed that Eliana asked her student (S5) to explain their ideas but there was no evidence, by applying the R-ASMM, that she encouraged the class to comment on the student (S5)'s explanation, for example, by creating or using assessment criteria to evaluate a student's models. This example corresponds to a novice level of proficiency for the criterion related to the construction and use of assessment criteria for the theoretical dimension of knowledge of peer and self-assessment. Eliana, before and after the OPDC, displayed a *novice* level of proficiency for the criterion related to the construction and use of assessment criteria to engage students in peer and self-assessment. At a novice level, an IST can encourage students to generate an explanation, but their peers are not challenged to develop or use assessment criteria to engage students in peer assessment by commenting on their classmates' ideas/models. In other words, in a novice level ISTs do not act to engage students in peer and self-assessment. The R-ASMM revealed that Eliana asked her students to explain his/her model to his/her peers on average once per class observed (1 time before the OPDC, and 1 time after the OPDC in one class) but she did not encourage students in the class to comment on the student explanation, for instance, by using assessment criteria to evaluate good models. Eliana asked a student in the fifth class before the OPDC, to explain to her classmates what an electron cloud is to clarify another students' question (see the excerpt from class observation in Fig. 16). The information from the class observation was complemented with the results in the interview which showed that before and after the OPDC Eliana kept the idea that students are not prepared to judge and assess their peers (see underlined portion from the interview data in the right-side of Figure 16). She did however expect them to explain the idea to others. By using the information from the classroom observations and interviews based on the analysis of the R-ASMM, Eliana's pedagogy for the

indicator related to the construction of assessment criteria was scored as a *novice* before and after the OPDC since no changes were identified in the class observations and interviews before and after the OPDC.

In a second example, to show how the R-LPAL was used to identify how ITSs' assessment literacy about models and models influenced their pedagogy, I use Samantha's case to illustrate her level of proficiency before and after the OPDC. Samantha initially showed an *advanced beginner* level of proficiency for one of the indicators used in the dimension of knowledge of peer and self-assessment; "Advanced assessor (Evaluate new information): Asks students to analyze new information to promote the evaluation and modification of models that help them collect evidence to show the utility and explanatory and predictive power of their generated models." In an advanced beginner level of proficiency, ISTs "Ask students to generate a model or provide an explanation and the IST asks them to review information without challenging them to evaluate their models." Based on the class observations, the R-ASMM coded that Samantha did not engage students in peer or self-assessment to evaluate a model before the OPDC. Nevertheless, the analysis of the interview before the OPDC showed that Samantha was able to reflect on the importance of challenging students to analyze and evaluate evidence or information to understand a phenomenon. For example, in the interview before the OPDC, she stated that after teaching reactivity of alkali metals, she often asks her students to solve exercises to determine the Gibbs free energy and evaluate the spontaneity of a process in a galvanic cell. She stated that the students, "Can begin to infer whether or not some reactions occur due to a trend that can be identified in the periodic table, as in the case of halogens" and "They [students] can evaluate whether a reaction is spontaneous based on the [chemical] potential that is

generated." Therefore, even though I did not observe the engagement of students in peer or self-assessment in Samantha's classroom, I scored her performance as an *advanced beginner* since she understood the relevance of challenging students to evaluate information or variables to justify their claims as suggested by her interview. It is worth noting that I did not score her knowledge of peer and self- assessment for this criterion as *novice* because the lowest level of proficiency referred to the fact that ISTs only "shows a model and explains the variables included, for example, in a diagram. S/he provides information that is used to understand a model and uses this information to evaluate the utility of a model instead of challenging students to do it." In Samantha's answers in the interview before the OPDC it can be noted that she suggested that students are in charge of generating the relationships and evaluating them after analyzing information form the periodic table and by determining and interpreting the Gibbs free energy. Surprisingly, Samantha, in the first class after the OPDC, implemented different assessment strategies that encouraged students to evaluate information and their models of the periodic table, such as i) asking students to analyze data (e. g., atomic radius) to evaluate a generated model of the periodic table; ii) asking students to compare their initial models of the periodic table; iii) analyze analogue data to identify exceptions of the periodic law in terms of the ordering of elements, and iv) encouraged her students to modify their initial relationships for the generated model of the periodic table. For instance, she said "Look at the table that says ionization. Now, see how the ordering that you made (for the model of the periodic table) responds to you. Are they in the same order?", "By electronegativity, where would we have placed the hydrogen?", "What if I ask you to use the periodic table to identify the position of an element that is $5p^6$". In this last example, Samantha guided her students about how to use their models of the periodic table and the information from the electron configuration to make a

prediction and identify the period of an element based on the energy level (5) and the group based on the number of electrons in the orbital (6). She also stated in the interview after the OPDC that she challenged her students to provide arguments to justify their claims and evaluate their models to "generate changes (in their initial models of the periodic table) in order to be able to reach the final (target) model by themselves." Based on these changes observed in Samantha's pedagogy and on her reflections during the interview after the OPDC, Samantha's level of proficiency for the criterion related to evaluate new information was scored as an *advanced assessor* after the OPDC. In other words, she "asked students to analyze new information to promote the evaluation and modification of models that help them collect evidence to show the utility and explanatory and predictive power of their generated models," consistent with the advanced assessor category in the R-LPAL. The R-LPAL levels for each of the ISTS on each of the dimensions in Figure 1 are detailed in the results chapter.

3.2.1.9 Triangulation

Based on a sequential mixed method design (Creswell et al., 2003) in which a combination of quantitative and qualitative data collection techniques were employed, triangulation and integration of findings were necessary (Moran et al., 2006). The integration of findings is useful to explore how findings obtained from quantitative and qualitative resources interact and converge in a study (Moran et al., 2006). In order to address the research questions in this study i) whether ISTs' understanding about models and modeling is related to ISTs' assessment literacy in MBT, and ii) how ISTs' assessment literacy about models and modeling changed as a result of attending an online training course for professional development in MBT, different research techniques were used to gather evidence and collect data. The data set was a

questionnaire, interviews, class observation videos and teachers' artifacts (e. g., exams, lesson plans). The data set was analyzed considering four levels of information (Contreras, 2010):

- Response statement level: This level involved ISTs' self-reported understanding of MBT and how to assess students by models and modeling based on the administration of a Likert-scale questionnaire (QALMBT questionnaire).
- Declarative level: This level provided information on ISTs' thinking and reflection about their own practices in MBT and offered data to respond to how ISTs' assessment literacy about models and modeling influenced their pedagogy and how their pedagogy changed as a result of attending an online professional development course on MBT. The instruments to collect data included two protocols for semi-structured interview.
- Planning and design level: This level included the lesson plans that ISTs developed before teaching their students. The lesson plans contributed to identifying aspects such as the goals and outcomes that teachers expect to achieve in their classes and reflect how ISTs evaluate their students in MBT.
- Practice/Enactment level: This level described the teaching practices that ISTs implemented in their pedagogy after planning their lessons. In other words, this level reflected the coherence between lesson planning and instruction. Specifically, I adapted Carlson and Daehler's (2019, p. 53) definition of enactment as "[T]he specific knowledge and skills utilized by an individual teacher in a particular setting, with [a] particular student or group of students, with a goal for those students to learn a particular concept [model] [during the generation, evaluation and modification of generated models]." ISTs' pedagogy was analyzed through the application of rubrics of observations and the analysis of teachers' instructional decisions and teaching strategies during instruction.

During the analysis and interpretation of the results, I used methodological triangulation to obtain a better understanding of science teachers' reality. This type of triangulation involves the combination of different methods to gather data in order to reduce bias and comprehend the phenomenon of interest (Abdalla et al., 2018). In my study, I used quantitative and qualitative methods to gather data which included the administration of a questionnaire, class observations, and semi-structured interviews. Different data sources were analyzed (QALMBT questionnaire, class observation Rubrics "R-ASMM", ISTs' artifacts such as exams and lesson plans) to answer the research questions and include credibility and trustworthiness to the study results. Each of the four levels of information (response statement level, declarative level, planning and design level, and practice/enactment level,) indicated above was contrasted and analyzed to explore ISTs' ALMBT. For example, the response level related to the administration of the QALMBT questionnaire was particularly valuable in addressing the first research question related to whether ISTs' knowledge of models and modeling related to ISTs' assessment literacy in MBT. This level also informed (through the exploratory factor analysis) the factors that were identified during the analysis of qualitative data from interviews (declarative level), observational rubrics (R-ASMM) and class observations (practice/enactment level). Also, the information obtained from interviews (declarative level) was compared with the practice/enactment level to compare the coherence between teacher ideas about assessment in MBT and how they put their knowledge into practice when assessing students' models. Each of the themes initially identified in the EFA and then confirmed during the analysis of qualitative data were also revised and enriched by analyzing information from the planning and design level such as teachers' administered exams and lesson plans. Hence, each data source was particularly valuable in

addressing research question 2 related to in what ways ISTs' assessment literacy about models and modeling influenced their pedagogy. Finally, the analysis of each of these levels during the triangulation of data helped to verify and confirm the findings of the study and support the accuracy of themes by comparing them with the factors identified initially in the quantitative phase.

3.2.2 Issues

3.2.2.1 Subjectivity Statement

I am an international student, and I worked in Chile as a secondary science teacher for three years. Prior to beginning my doctoral program, I worked in a public university as a teaching assistant for the subject Natural Sciences Didactics and as a practicum coordinator for the Chemistry and Biology teacher program for five years. In this role, I supervised and guided science teacher candidates throughout their practicum, which helped me to better understand their pedagogy. In this vein, the rubrics and the data collection procedures in this study have been undoubtedly influenced by my own experience as a practicum coordinator. I must acknowledge that my experience as a graduate student in a Canadian university plays a key role in the way I designed the study and how I analyzed and interpreted the data. I attended courses in research in the teaching and learning of the sciences and measurement and assessment which have impacted the presumptions and lens that I have used to conduct the study and analyze the data. For example, the knowledge about MBT that I have acquired from the revision of the literature in MBT from previous research may have influenced how I analyzed ISTs' pedagogy. I have been conscious about registering not only ISTs' teaching strategies that are framed in an

MBT approach but also, I analyzed those student-centered strategies that were developed by ISTs.

3.2.2.2 Trustworthiness

Trustworthiness is used to identify potential bias in a study. Criteria such as credibility, transferability, dependability, and confirmability are often included in studies to support qualitative research (Anney, 2014; Lincoln & Guba, 1986; Schwandt et al., 2007; Shenton, 2004). From a positivistic perspective, criteria to enhance trustworthiness include internal validity, external validity/generalisability, reliability, and objectivity, respectively (Guba, 1981). In the case of the quantitative data collected by administering the QALMBT questionnaire, I report reliability and validity to ensure the trustworthiness of the data. Trustworthiness in my interpretive study was sought by addressing these criteria as elaborated below.

3.2.2.3 Validity and Reliability

According to Hubley and Zumbo (2011), "validity is about whether the inference one makes is appropriate, meaningful, and useful given the individual or sample with which one is dealing and the context in which the test user and individual/sample are working" (p. 220-221). Three common forms of validity include content, construct, and criterion validity (Shrotryia & Dhanda, 2019). Content and construct validity (by using Exploratory Factor Analysis) were employed in this study as evidence to validate the QALMBT questionnaire.

Content validity is evidence of validity based on subject matter experts' judgments. It refers to "[T]he degree to which a sample of items, taken together, constitute an adequate operational

definition of a construct" (Polit & Beck, 2006, p. 490). In my study, content validity was used to evaluate relevance, pertinence, and consistency of items in the QALMBT questionnaire in terms of the content (Schwartz & Barbera, 2014). Firstly, for the English version, the questionnaire was reviewed by five subject matter experts in the area of science education. The subject matter experts were a science teacher from B.C., a Ph.D. student in science education, and three scholars in science, science education, or education (Ph.D.). Each of the experts were requested to review the appropriateness of the items in terms of the eight dimensions included in the framework for assessment literacy (Figure 1). Overall, the subject matter experts agreed with the items from the questionnaire. Most of the expert comments were related to the accuracy or semantics of the language. For example, initially, one of the items stated, "In my classes I ask students to criticize the models created by their classmates" was modified later because two of the experts pointed out that the word "criticize" in a Canadian context could potentially elicit negative thoughts. After the feedback, the item was reformulated as "In my classes, I ask students to comment on the models created by their classmates."

It is worth mentioning that one common approach to determine content validity is by calculating the content validity index (CVI), which corresponds to the proportion of content experts who consider each scale item as relevant. Lynn (1986) suggests that in a panel of "[F]ive or fewer experts, all must agree on the content validity for their rating to be considered a reasonable representation of the universe of possible ratings" (p. 383) in order to reduce chance factors. Because Lynn's suggestion was too conservative, and due to the fact that five experts rated the English version of the questionnaire, in my study, if an item had a degree of agreement lower than .4 in both version of the questionnaire (QALMBT-Generic and QALMBT-Modeling), the

item was dropped from the questionnaire. This situation only happened with the item "I use assessment to evaluate students' creativity when developing a model." When an item had a degree of agreement equal to 1 in one of both versions and no lower than .8 in the other version, the item was kept in the questionnaire and was revised based on subject matter experts' comments. Items with 0.6 of agreement were carefully revised and discussed with one of the experts who specializes in MBT until agreement was reached. Only one item with an agreement of 0.4 in the QALMBT-Generic and 0.6 in the QALMBT-Modeling was kept. This item asked, "For those areas that students have difficulty in comprehending, I promote the generation of a consensus explanation that helps students have a similar understanding of the core ideas." The item was revised and kept in the final version of the questionnaire because the development of consensus models and explanations is widely studied in a number of studies related to model-based teaching and learning (Cheng & Brown, 2015; Gilbert, 2004; Gilbert & Justi, 2016; Justi, 2000).

To explore the degree of consensus among subject matter experts, a supplement to CVI is the kappa coefficient ($\varkappa$). This index is also suggested as a supplement to CVI because its formula adjusts for chance agreement (Honda & Ohyama, 2020). Kappa is a consensus index of interrater agreement or also referred as interrater reliability (IRR) (McHugh, 2012); however, in some cases, Kappa provides low values even though there is a high percentage of agreement. This is referred as the "kappa paradox" (Warrens, 2010; Zec et al., 2017). An alternative to Cohens' kappa statistic is suggested by the computation of the Agreement Coefficient 1 (AC1) indicated by Gwet (2001), which is more robust in terms of this paradox (Honda & Ohyama, 2020; Zec et al., 2017). Gwet's AC1 is a liberal estimate of reliability (Zhao at al., 2013) and is a chance-

adjusted method that, "Uses average rather than individual marginal totals in the calculation of chance probabilities" (Grant, Button & Snook, 2017, p.3; Gwet, 2001; Zec et al., 2017). Gwet's AC1 ranges from -1 to 1, and a three-degree scale is often used for the interpretation of the agreement coefficient (Karstad et al., 2018). The benchmark scale includes lower than .40 as poor, between .40 to .75 as intermediate to good and more than .75 as excellent (Fleiss, Levin & Paik, 2003). Values of AC1 ≤ 0 are classified as no agreement (Wongpakaran et al., 2013). In my study, I used Gwet's AC1 as a measure of IRR. For the English version of the QALMBT, five subject matter experts were asked to rate each item, whether it was essential to measure the dimension or sub construct. Gwet's AC1 was preferred in comparison to Cohen's kappa because the values for Gwet's AC1 were similar numbers to the percentage of agreement whereas Cohen's kappa was surprisingly low due to "kappa paradox". To calculate IRR, unweighted Gwet's AC1 was computed independently for each coder pair who assessed the items using a 2-level ordinal scale (0 = no essential; 1 = essential). Then, an average was calculated to obtain a single index of IRR. The resulting index suggested a Gwet's AC1 index of .68 for the QALMBT-General and a percentage of agreement of .86. For the QALMBT-Modeling version, Gwet's AC1 was .81, and the percentage of agreement was .90. Finally, for the QALMBT-Epistemic, Gwet's AC1 gave a score of .93 and a percentage of agreement of .97. For the Spanish version of the questionnaire, items were translated to Spanish, and two experts participated in this phase (The process of back-translation was detailed in the research design and methods chapter). The values for Gwet's AC1 for each section of the questionnaire were QALMBT-Generic (.84), QALMBT-Modeling (.52), and QALMBT-Epistemic (.95) (see appendix H for more details about subject matter experts' ratings on the QALMBT). In the case of the QALMBT-Modeling for the Spanish version, the instrument showed a low inter-rater

reliability because only two experts rated the questionnaire. Regarding the qualitative data, it is worth mentioning that internal validity of the data analysis was also pursued during the qualitative coding of the data. Six cycles of independent revisions of the codes were conducted with the academic supervisor for excerpts of the data until reaching an agreement greater than 90%. Disagreements were discussed and resolved.

3.2.2.4 Score Reliability of the Scale

Internal consistency of the QALMBT questionnaire, which is related to the homogeneity of the items or the extent to which a group of items measure a construct (Henson, 2001), was assessed by calculating Cronbach's alpha and omega coefficient. In science education, an arbitrary value of .70 is often used as acceptable reliability for a scale (Taber, 2018b). Cronbach's alpha is often used to calculate the reliability of a scale even though it is not considered to be on its own an adequate measure for estimating the internal consistency of data that are not continuous. Nevertheless, it is still often used to report the internal consistency of 5-point Likert format scales that can be treated as continuous. The QALMBT questionnaire included an ordinal categorical scale in a Likert format. I reported Cronbach's alpha "α", but I also determined an alternative measure to assess the homogeneity of the scale for data that are not continuous- the omega coefficient "ω" (McDonald, 1999). Cronbach's alpha and the Omega coefficient were calculated using the R package (psych). For the Canadian version of the questionnaire, the values for Cronbach's alpha for the QALMBT-Generic, QALMBT-Modeling, and QALMBT-Epistemic were .9, .97, and .87, respectively. For the omega coefficient, the values were respectively .93, .97, and .91. The Chilean version of the questionnaire showed a similar trend to the Canadian instrument. The Cronbach's alpha was .94 for the QALMBT-Generic, .96 for the

QALMBT-Modeling, and .86 for the QALMBT-Epistemic. The Omega coefficient for each section of the QALMBT was .95, .96, and .89, respectively. In other words, the QALMBT questionnaire had relatively high internal consistency since the alpha and omega coefficient were higher than the cut-off of .70 suggested by the literature.

3.2.2.5 Credibility

According to Morrow (2005), researchers can achieve credibility, for example, by prolonged engagement with participants, conducting observations in the natural setting, and implementing a peer researcher review. In my study, I created the instruments based on the literature review and frequent debriefing sessions with peer researchers. These instruments were reviewed by a group of subject matter experts to ensure they measured what was intended. Credibility was pursued by conducting a comparative analysis in which an iterative process of coding and triangulation of data from different data sources, such as interview transcripts, class observation, and teachers' and students' artifacts, was considered. Also, peer debriefing of coding and themes through meetings with peer researchers contributed to the credibility of this study.

In qualitative research, the researcher's involvement in the study may potentially influence the prepositions and conclusions of the study. This subjectivity was minimized in the process of data analysis by using thematic analysis. Specifically, axial coding was used as a qualitative research technique to construct linkages between data to ensure the trustworthiness and credibility of the claims by comparing and contrasting pre-established categories and subcategories with emergent themes that appeared from the data. This process of coding analysis involved inductive and deductive reasoning to explore the relationship between codes (Allen, 2017). To ensure

credibility and transparency in the study, participants were given the right to withdraw from the study at any time. During the process of data collection, the five ISTs who participated in the study continued in the study; however, the overall number of class observations depended on the context of each school and teachers' availability. Unfortunately, two ISTs had difficulties in conducting regular teaching sessions due to midterm exams, which reduced the number of classes that could be observed. The political context that affected the country from mid-October, 2020 affected the capacity to observe as well. Similarly, one teacher who was observed five times during the first cycle of class observation (James), was only observed three times in the second cycle of class observations because the school year finished earlier. It is worth noting that theoretical saturation was achieved in each case in the cycle of class observations since similar assessment practices were identified over and over again and no additional data were found. All participants had expressed their interest and availability to participate in the whole process, but external socio-political factors that were beyond my control influenced this process somewhat.

3.2.2.6 Transferability

This criterion corresponds to the extent of generalizing the results of a study to different contexts. Transferability can be achieved when the researcher offers enough information about the instruments, context, processes, and participants to facilitate the transferability of the study (Morrow, 2005). The transferability of the results in this study was sought instead of generalizability across contexts by offering a cross-case analysis of five ISTs. Although this sample is small, I provided detailed information about participants based on the gathering of data from different resources such as interviews, class observation and teachers' artifacts. A thick description of contextual and experiential information was further provided to facilitate the

understanding of participants' behavior in different events and actions (Dawson, 2012) in the science classroom and suggest how concepts from the study might transfer to similar contexts.

3.2.2.7 Dependability

This criterion refers to the consistency of a study across time, researchers, and analysis techniques (Gasson, 2004). In this study, all the instruments and analysis techniques that were used to collect and analyze the data are provided in the appendices. Morrow (2005), for example, suggests that researchers must furthermore track a detailed chronology of research activities and processes that influence i) how the data was collected and analyzed and ii) how the themes, categories or models emerged. I provided detailed description of how the instruments were created, validated, and administered throughout the research. An audit trail was conducted as a strategy to establish that the findings emerged from the data and not from the researcher's preconceptions alone (Cutcliffe & McKenna, 2004). An audit was performed by the academic supervisor on each phase of the research and data analysis to strengthen the research.

3.2.2.8 Ethical Procedures

A consent form was provided in advance to the participants via email in order to give them at least a week to decide on their participation in the study. Once ISTs received the consent form and agreed to participate, they immediately received, via email, a link with access to the questionnaire. The consent form was also presented in the questionnaire and participants who agreed to participate were asked to select a checkbox for accepting to participate in the research. In the case of Chile, once participants answered the questionnaire, ISTs from subsidized and private schools from the metropolitan area of the region were informed about the second phase

of the study and were invited to participate in the study. An information sheet for parents/guardians was given to students which explained the purpose, risks, and scope of the study. Before starting the class observations, I attended each class and explained the purpose of the study. The participants were informed that I was only interested in observing the teacher-student interaction in the class, and that I was not going to assess or grade students' performance.

Chapter 4: Results and Discussions

In this chapter, I present and discuss how ISTs' knowledge of the nature of models and modeling influences their assessment literacy and how assessment literacy and assessment practices might change with professional development. Specifically, I attempt to answer the two research questions of the study related to:

> **Research Question 1:** Are ISTs' knowledge of models and modeling related to ISTs' assessment literacy in MBT?
>
> **Research Question 2:** In what ways do ISTs' assessment literacy about models and modeling influence their pedagogy?

It is worth mentioning that in this study, assessment literacy in MBT (ALMBT) assumes that teachers' knowledge of models and modeling might be connected to the way that teachers design and implement assessment strategies in the science classroom. Also, I assume that teachers might show different levels of assessment literacy, which given the first assumption, depends on their level of knowledge of the role of models and modeling. Moreover, I hypothesize that teachers might show different levels of ALMBT since assessment literacy in this study is conceptualized as knowledge that inform teachers' assessment practices. In order to respond to research question 1, I do not only report an item-level analysis on ISTs' knowledge of models and modeling and their assessment practices when engaging students in thinking with models, but I also detail the number of factors identified in the exploratory factor analysis (EFA). The factors identified in the EFA and conceptualized based on Fig. 1 were used to make inferences about ISTs' patterns of responses. In other words, based on observed variables (items), ISTs' responses were grouped to uncover groups of assessment practices that they reported they implemented more or less often

when assessing students' reasoning with models. Then, I present regression models to identify the relationship between ISTs' knowledge of the nature of models and modeling and their assessment literacy in MBT.

I also detail the emergent themes obtained from an analysis of the ISTs' assessment literacy before and after an online professional development course. The factors identified in the baseline phase of ISTs' assessment literacy in MBT were used as themes to inform the qualitative analysis. Also, I want to emphasize that the illustrative examples included in the results are meant to both exemplify teachers' assessment repertoire and problematize ISTs' assessment literacy in MBT. As stated in chapter 3, these examples are not intended to suggest a generalization of the results.

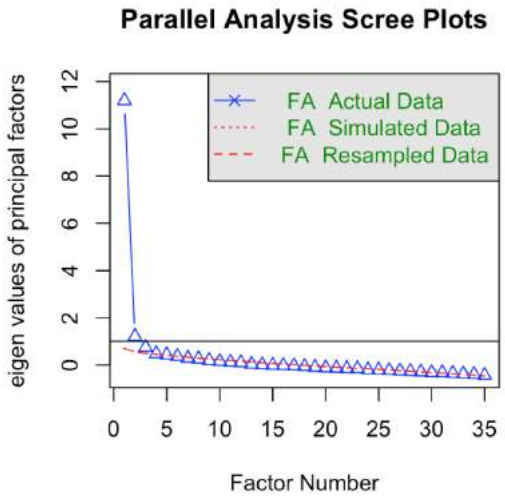
4.1 Results of the Exploratory Factor Analysis and Conceptualization of the Extracted Factors for each Section of the QALMBT

Three hundred and eighty-six Chilean ISTs completed the QALMBT questionnaire online. As shared previously, the first section included a 35-item Likert-type scale in which respondents indicated how often they implemented specific assessment strategies based on a) a general approach to teaching science and, b) when teaching science with models. The scale included five points; never (1), very rare (2), sometimes (3), frequently (4), and very frequently (5) and ISTs scored a specific common statement related to their assessment practices when teaching science. The second section of the questionnaire, which is presented later, included 20 items that measured ISTs' knowledge of the nature of models and modeling.

Exploratory factor analysis (EFA) models were estimated for each component of the QALMBT questionnaire and Maximum likelihood (ML) estimation was used with direct oblimin rotation to conduct the EFA. The first step before conducting the EFA included the identification of the number of factors or components based on the analysis of scree plots as a visual depiction for the eigenvalues. I conducted this procedure in parallel by the QALMBT-Generic, QALMBT-Modeling and QALMBT-Epistemic and the results of the EFA are reported in the same order. Figure 17 shows the scree plot with a visual depiction of the eigenvalues for the QALMBT-Generic to identify the number of factors related to assessment literacy in a general approach of teaching. Three factors were identified after exploring the scree plots and parallel analysis. Based on the Kaiser criterion at eigenvalues larger than 1 or .7, the results suggested a two-factor and a three-factor model, respectively. Based on the ratio of the first-to-second eigenvalues, only one factor was suggested. EFAs with a three, two, and one-dimensional models were checked and compared.

Figure 17

Scree Plot for the QALMBT-Generic



After removing one item at the time for the 3-factor model whose loadings were smaller than 0.3 or split across several factors, the results of the EFA revealed that the third factor did not include any items loading larger than .3; therefore, a new model was explored with only two factors. The factor loadings of the item for the two-factor model lacked theoretical interpretation, and a model with only one factor was explored. The 1-factor model had an acceptable fit with a RMSEA (Root Mean Square Error of Approximation) at .066, 90% CI [.06, .068], and RMSR (Root Mean Square Residual) with acceptable fit (0.06), poor CFI (Comparative Fit Index) (.820) and poor TLI (Tucker-Lewis Index) (.809). The reliability of the factor of .94 was excellent (M = 4.2; SD = 0.5). An examination of the goodness-of-fit information for the EFA results showed that a 1-factor model better represented the dimensionality of the QALMBT-Generic. Factor 1 included each of the items from the QALMBT and the reliability of the factor was .94 (M = 4.2; SD = .5). Table 15 shows the factor loadings for each item and Table 16 summarizes the results for each EFA based on the criteria used to determine the number of factors. In the case of the data without outliers, a 3-factor model had a good fit with a RMSEA that indicated good fit at .044, 90% CI [.033, .052], and RMSR with good fit (.03), CFI (.960), and TLI (.945).

Table 15

Item Factor Loadings for each Factor Solution for the QALMBT-Generic

Data with outliers							
Item	Factor 1	Item	Factor 1	Item	Factor 1		
A1	**0.42**	A11	**0.47**	A21	**0.62**	A31	**0.70**
A2	**0.46**	A12	**0.56**	A22	**0.56**	A32	**0.56**
A3	**0.41**	A13	**0.54**	A23	**0.47**	A33	**0.71**
A4	**0.42**	A14	**0.56**	A24	**0.66**	A34	**0.68**
A5	**0.50**	A15	**0.47**	A25	**0.62**	A35	**0.54**
A6	**0.50**	A16	**0.56**	A26	**0.68**		
A7	**0.56**	A17	**0.59**	A27	**0.67**		
A8	**0.53**	A18	**0.60**	A28	**0.47**		
A9	**0.53**	A19	**0.55**	A29	**0.67**		
A10	**0.56**	A20	**0.45**	A30	**0.69**		

Note: Factor loadings in bold type were considered for the conceptual interpretation. The factor loadings for the data without outliers are provided in appendix F.

Table 16

Summary of Model Fit for each Model for the QALMBT-Generic

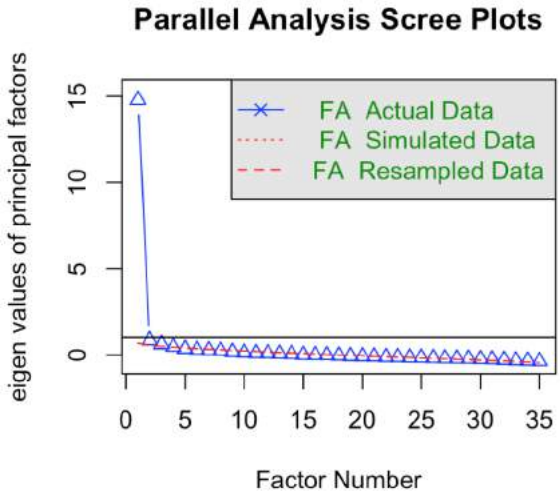
Model	Items deleted	CFI	RMSEA(90%CI)	TLI	RMSR	Reliability of factors(Cronbach's alpha; *M; SD)*
Based on the Parallel Analysis						
3- factor model**	1,4,6,10,11,13,16,19,20,22,23,26, 27	.969	.039[.028, .046]	.957	.03	F1 = .89; 4.2; .63 F2 = .80; 4,2; .59 F3 = .62; 4.5; .63
Based on eigenvalues (eig.> 1)						
2-factor model***	16	.888	.055[.048, .057]	.872	.05	F1 = .88; 4.1; .55 F2 = .91; 4.3; .60
Based on the ratio of the first-to-second eigenvalues						
1-factor model*		**.820**	**.066[.06, .068]**	**.809**	**.06**	**F1 = .94; 4.2; .54**
Based on data without outliers						
3-factor model	2, 3, 7, 8, 9, 10, 15, 16, 18, 19, 21, 22, 23, 26	.960	.044 [.033, .052]	.945	.03	F1 = .89; 4.3; .56 F2 = .71; 3.9; .73 F3 = .71; 4.1; .61

Note: M = Mean; *SD* = Standard deviation; CFI= Comparative Fix Index; RMSEA= Root Mean Square Error of Approximation; TLI= Tucker Lewis Index; RMSR= Root Mean Square of Residuals. * Model preferred to represent the dimensionality of the questionnaire (bold type). ** Last factor only included two items; therefore, the factor was dropped, and the model was checked again with only two factors. ***A new factor model was run with only one factor because the factors lacked theoretical interpretation.

Because I assume that the QALMBT-Generic and QALMBT-Modeling measured ISTS' assessment literacy from two different approaches, general and MBT, I conducted the EFAs separately. Figure 18 provides a visual depiction of the eigenvalues for the QALMBT-Modeling. The scree plot and parallel analysis revealed 3 factors, whereas the Kaiser criterion at eigenvalues larger than 1 and .7 only showed 1 and 2 factors, respectively. The ratio of the first-to-second eigenvalues suggested only 1 factor.

Figure 18

Scree Plot for the QALMBT-Modeling



Models with 1, 2, and 3 factors were explored and compared. The CFI, TLI, RMSEA and RMST showed that by decreasing the number of factors, the model fit worsens. Therefore, a 3-factor model was preferred to better fit the data. This model was tested after deleting items 2, 3, 6, 18, 23 and 27 (because they loaded in two or more factors or because the factor-loading was smaller than .3) and showed a good fit for each of the measures for model fit (CFI = .949; RMSEA = .005 [90% CI = .042, .054]; TLI= .936; RMSR = .03). Table 17 shows the factor loadings for the data set. Table 18 summarizes the values for the model fit obtained for each model tested. Factor

1 included nineteen items that measured “Implementation of strategies to promote the elicitation and assessment of students’ models.” Examples of items of this factor included “19. After a summative assessment, I clarify common students’ misconceptions or alternative ideas about their generated models in class,” “22. When students express their claims in front of the classroom, I establish classroom norms to promote a safe expression of students’ ideas about their models,” and “29. I challenge my students to show evidence to support their claims about their models,” Factor 2 included seven items that measured “Self and peer assessment of generated models.” For example, item 4 asked “I challenge my students to develop assessment criteria to evaluate the models constructed by their classmates,” whereas item 11 asked “13. In my classes I ask students to comment on the models created by their classmates.” Finally, factor 3 included three items that appear to assess “Communication of assessment criteria to assess students’ models.” Examples of these items include “5. I explain to students the criteria that I will use to evaluate their models,” and “7. I use results from an assessment to compare how students' ideas about a model have been reshaped.” The reliability of each factor was excellent for factor 1 (M = 3.8, SD= .72), good for factor 2 (M = 3.6, SD = .8) and acceptable for factor 3 (M = 4.0, SD = .88).

Table 17

Item Factor Loadings for each Factor Solution for the QALMBT-Modeling

Data with outliers				Data without eliminating items			
Item	Factor 1	Factor 2	Factor 3	Item	Factor 1	Factor 2	Factor 3
B9	**0.33**	0.22	0.19	B9	0.28	**0.30**	0.18
B12	**0.31**	0.08	0.27	B12	**0.13**	0.09	0.23
B15	**0.35**	0.25	0.21	B15	**0.32**	**0.30**	0.21
B16	**0.38**	0.11	0.27	B16	**0.41**	0.09	0.25
B19	**0.42**	0.00	0.15	B19	**0.43**	-0.03	0.16
B20	**0.34**	0.26	0.07	B20	**0.32**	0.27	0.07
B21	**0.65**	-0.01	0.17	B21	**0.62**	0.03	0.17
B22	**0.55**	-0.15	0.17	B22	**0.59**	-0.19	0.19
B24	**0.58**	0.12	0.04	B24	**0.59**	0.12	0.03
B25	**0.68**	-0.04	0.04	B25	**0.66**	0.01	0.03
B26	**0.68**	-0.01	0.17	B26	**0.67**	0.00	0.20
B28	**0.74**	-0.12	-0.06	B28	**0.73**	-0.07	-0.09
B29	**0.65**	0.24	-0.11	B29	**0.61**	**0.31**	-0.13
B30	**0.69**	0.09	-0.07	B30	**0.67**	0.10	-0.08
B31	**0.65**	0.18	-0.05	B31	**0.63**	0.18	-0.03
B32	**0.65**	-0.05	0.02	B32	**0.66**	-0.06	0.02
B33	**0.73**	0.07	-0.02	B33	**0.74**	0.03	0.01
B34	**0.79**	-0.07	0.00	B34	**0.78**	-0.08	0.04
B35	**0.46**	0.17	0.06	B35	**0.46**	0.14	0.08
B1	0.02	**0.38**	0.07	B1	0.02	**0.39**	0.05
B4	-0.07	**0.76**	0.03	B4	-0.08	**0.72**	0.06
B10	0.27	**0.30**	0.23	B10	0.29	0.27	0.24
B11	0.23	**0.47**	0.05	B11	0.19	**0.53**	0.03
B13	0.21	**0.52**	0.05	B13	0.22	**0.49**	0.07
B14	0.24	**0.39**	0.22	B14	0.25	**0.36**	0.21
B17	0.13	**0.61**	0.14	B17	0.12	**0.64**	0.11
B5	0.02	-0.01	**0.74**	B5	0.04	0.00	**0.74**
B7	0.21	0.22	**0.34**	B7	0.20	0.27	**0.31**
B8	0.00	0.08	**0.80**	B8	0.04	0.11	**0.73**
				B2	0.08	0.26	0.15
				B3	0.18	**0.30**	0.06
				B6	-0.05	**0.49**	**0.34**
				B18	**0.32**	**0.39**	0.21
				B23	**0.30**	0.10	0.34
				B27	**0.45**	**0.44**	-0.10

Note: Factor loadings in bold type were considered for the conceptual interpretation.

Table 18

Summary of Values for Model Fit for each Model for the QALMBT-Modeling

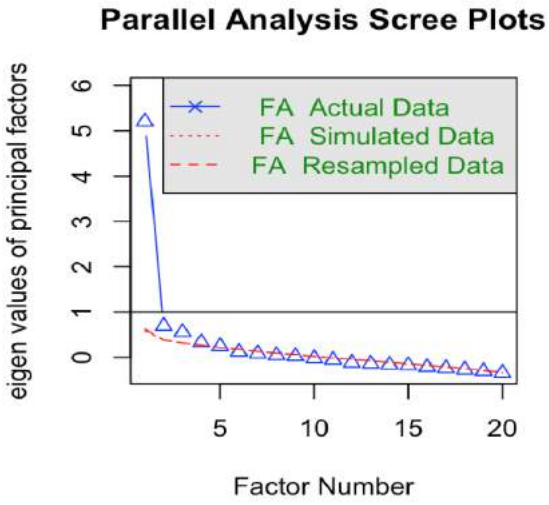
Model	Items deleted	CFI	RMSEA(90%CI)	TLI	RMSR	Reliability of factors(Cronbach's alpha; *M; SD)*
Based on the Parallel Analysis						
3- factor model*	**2, 3, 6, 18, 23, 27**	**.949**	**.05[.042, .054]**	**.936**	**.03**	**F1 = .94; 3.8; .72** **F2 = .85; 3.6; .80** **F3 = .78; 4.0; .88**
Based on eigenvalues > .7						
2-factor model	1,2,3,4,7,9, 10,12,14,17, 18, 23	.944	.057[.049, .063]	.932	.04	F1 = .94; 3.8; .73 F2 = .78; 3.9; .91
Based on the ratio of the first-to-second eigenvalues and eigenvalues > 1						
1-factor model		.885	.063[.057, .065]	.877	.05	F1 = .96; 3.8; .69
Based on data without outliers						
3 factor model	2, 3, 6, 9, 11, 14, 16, 23	.964	.046 [.037, .051]	.954	.03	F1 = .95; 3.8; .73 F2 = .8 ; 4;.85 F3 = .76; 3.5; .83

Note: M = Mean; *SD* = Standard deviation; CFI= Comparative Fix Index; RMSEA= Root Mean Square Error of Approximation; TLI= Tucker Lewis Index; RMSR= Root Mean Square of Residuals. * Model preferred to represent the dimensionality of the questionnaire (bold type)

Because the QALMBT-Epistemic measured ISTs' knowledge about models and modeling and not their assessment literacy, a separate EFA was conducted to identify the factors or components related to ISTs' epistemological knowledge. Figure 19 shows the scree plot for the QALMBT-Epistemic, which revealed 3 factors, whereas the parallel analysis suggested 4 factors. In the case of the analysis of eigenvalues greater than 1.0, .7, and the ratio of the first-to-second eigenvalues suggested only 1 factor. Each model was run, and the values for model fit were compared.

Figure 19

Scree Plot for the QALMBT-Epistemic



A 1-factor model showed poor fit to the data and did not allow the identification of sub-constructs included in the scale. A 4-factor model was run based on the parallel analysis. After reaching the simple structure, factor 4 showed low reliability (see Table 20), and a model with only three factors was explored to compare the results. After removing items based on loadings smaller than 0.3 and then after checking cross-loadings, the three-factor solution allowed the identification of three sub-dimensions related to the epistemology of models. Table 19 shows the

factor loadings for the factor solutions for the data before and after removing items. Even though some researchers conceptualize the epistemology of modeling as comprised of five (see, for example, Krell & Krüger, 2016) or four sub-dimensions (Oliva & Blanco-López, 2021), the results of a 3-factor model for the QALMBT-Epistemic summarized three important aspects related to i) "Generative tools for testing scientific knowledge", ii) "Tentative nature of models", and iii) "Multiplicity of scientific models". The three-factor model (CFI = .945; RMSEA= .047[90% CI = .035, .057]; TLI = .917; RMSR = .04) was consistent with the underlying theory guiding the design of the QALMBT .and showed a good fit of the data. The reliability of each factor was over .61; therefore, the reliability of the factors was sufficient for factor 1 (M = 4.2, SD= .58), factor 2 (M = 4.2, SD = .63) and factor 3 (M = 4.2, SD = .66). Table 19 summarizes and compares the goodness-of-fit information for models tested for the QALMBT-Epistemic.

Table 19

Item Factor Loadings for each Factor Solution for the QALMBT-Epistemic

Data with outliers				Data without eliminating items			
Item	Factor 1	Factor 2	Factor 3	Item	Factor 1	Factor 2	Factor 3
C2	**0.42**	0.15	-0.07	C2	**0.44**	0.14	-0.09
C8	**0.62**	0.04	-0.04	C8	**0.63**	0.05	-0.06
C9	**0.37**	0.24	0.06	C9	**0.37**	0.27	0.02
C10	**0.34**	-0.09	0.21	C10	**0.37**	-0.13	0.24
C12	**0.63**	-0.07	-0.03	C12	**0.61**	-0.05	-0.04
C13	**0.40**	0.15	-0.02	C13	**0.43**	0.09	0.03
C15	**0.55**	0.14	-0.01	C15	**0.58**	0.12	-0.03
C16	**0.52**	-0.02	0.25	C16	**0.53**	0.02	0.20
C18	**0.52**	0.01	0.12	C18	**0.54**	0.06	0.03
C1	-0.02	**0.42**	-0.03	C1	-0.01	**0.41**	-0.04
C3	-0.03	**0.65**	0.11	C3	-0.03	**0.66**	0.12
C4	0.17	**0.58**	-0.07	C4	0.17	**0.58**	-0.09
C5	0.11	**0.50**	-0.02	C5	0.10	**0.50**	-0.03
C11	0.05	**0.38**	0.22	C11	0.03	**0.40**	0.26
C6	-0.08	0.05	**0.48**	C6	-0.08	0.06	**0.52**
C17	0.06	0.23	**0.45**	C17	0.09	0.28	0.39
C19	0.01	0.28	**0.48**	C19	0.06	0.29	**0.44**
C20	0.19	-0.10	**0.57**	C20	0.21	-0.01	**0.47**
				C7	0.22	0.22	0.26
				C14	**0.43**	-0.14	**0.36**

Table 20

Summary of Values for Model Fit for each Model for the QALMBT-Epistemic

Model	Items deleted	CFI	RMSEA(90%CI)	TLI	RMSR	Reliability of factors(Cronbach's alpha; *M; SD)*
Based on the Parallel Analysis						
4-factor model**	2,7,14,15	.983	.032[.008, .046]	.967	.03	F1 = .68; 4.2;.63 F2 = .63; 4.2; .66 F3 = .67; 4.3; .68 F4 = .56; 3.9; .78
3-factor model*	**7, 14**	**.945**	**.047[.035, .057]**	**.917**	**.04**	**F1 = .69; 4.2; .58** **F2 = .68; 4.2; .63** **F3 = .63; 4.2; .66**
Based on the ratio of the first-to-second eigenvalues and eigenvalues .> 1; >.7						
1-factor model	1	.847	.068[.059, .075]	.828	.06	F1 = .87; 4.2; .51
Based on data without outliers						
3-factor model	10, 11, 13	.963	.041 [.025, .052]	.944	.04	F1 = .76; 4.3; .52 F2 = .66; 4.3; .56 F3 = .66; 4.2; .57

Note: M = Mean; *SD* = Standard deviation; CFI= Comparative Fix Index; RMSEA= Root Mean Square Error of Approximation; TLI= Tucker Lewis Index; RMSR= Root Mean Square of Residuals. * Model preferred to represent the dimensionality of the questionnaire (bold type). ** Factor 4 of this model showed low reliability; therefore, a model with only 3 factors was explore.

4.1.1 Conceptualizing QALMBT Emergent Factors

In this section, I conceptualize each of the factors extracted from the EFAs for each section of the QALMBT (Generic, Modeling, Epistemic). As it was already mentioned, the QALMBT-Generic measured ISTs' assessment literacy from a generic approach to teaching science centered on assessing disciplinary knowledge and core ideas in science without emphasizing the use of models. Only one factor was identified for the data set from the QALMBT-Generic with outliers. This factor was called "Assessment Literacy in Science Classroom" because it included each item from the QALMBT-Generic. It is worth mentioning the data set without outliers showed that a 3-factor model had a better fit to the data compared to the 1-factor model from the data with outliers. Even though I decided to keep the model with only one factor, I detail the 3-factor model only to show evidence that the results from EFA were different when including or removing outliers. Below, I conceptualize each factor for the data without outliers in order to inform the dimensionality of the results obtained from this analysis.

The first factor of the three encompassed many of the items included in each subcategory of the questionnaire. Table 21 indicates each of the items that were included in this factor. This factor was labeled *Intentions of teachers' assessment practices to promote the expression of students' ideas* as the questions covered ISTs' intentions and decisions to develop and implement assessment in the science classroom. The second factor included items that measure the extent to which teachers engage students in self and peer assessment of the core ideas in science. Table 22 shows the items included in this factor referred to as *Self and peer assessment of the understanding of scientific core ideas.* To the third factor were assigned items that mostly measure teachers' disciplinary knowledge and pedagogical content knowledge when teaching

inquiry in the science classroom, for example, by carrying out investigations, analyzing evidence and testing hypotheses. Therefore, the name of this factor was *Teacher assessment practices in inquiry*. The items included in this factor are detailed in Table 23.

Table 21

Detail of the Items Included in the Factor Related to Intentions of Teachers' Assessment Practices to Promote the Expression of Students' Ideas of the QALMBT-Generic for the Data Set Without Outliers

Item	Factor Loading	Theoretical dimension (Knowledge of…)	Statement
A31	0.66	Assessment purpose, content and methods	I use assessment to evaluate the internal consistency or coherence of students' ideas.
A27	0.51	Assessment purpose, content and methods	I assess how students make judgements in science based on reasoning.
A32	0.51	Grading	When I develop assessment instruments to assess students' understanding, I think beforehand how I will interpret the results.
A28	0.41	Grading	When I develop distractors (incorrect or inferior alternatives) in a test, I have in mind the different alternatives or inaccurate ideas that students might have for the content that they are studying.
A33	0.80	Feedback	I use assessment to give formative feedback to students about their understanding of the core ideas studied in class.
A34	0.73	Assessment interpretation and communication	I use assessment to locate evidence about the missing elements that students have not understood regarding the core ideas under study.
A24	0.65	Assessment interpretation and communication	I use the results generated from formative assessment to adjust the content of my lessons regarding the core ideas under study.
A29	0.55	Peer and self-assessment	I challenge my students to show evidence to support their claims about the content that they are studying.
A30	0.80	Assessment ethics	I use the results of the assessment to coach a student when she/he/they are having problems understanding a disciplinary core idea in science.
A25	0.56	Scaffolding and learning progression	I encourage students to use their pre-existing ideas in order to help them to construct an initial explanation that can be enriched later.
A12	0.34	Scaffolding and learning progression	When I make an attempt for students to understand a core idea in science, I organize the content in my lessons following a sequence which considers how student understanding can evolve over a span of time.

Note: The items were sorted based on the theoretical subcategory used to theorize assessment literacy.

Table 22

Detail of the Items Included in the Second Factor Related to Self and Peer Assessment of the Understanding of Scientific Core Ideas of the QALMBT-Generic for the Data Set Without Outliers

Item	Factor Loading	Theoretical dimension (Knowledge of…)	Statement
A17	0.62	Peer and self-assessment	I teach students how to judge the quality of their explanations based on the consistency of their ideas.
A4	0.59	Peer and self-assessment	I challenge my students to develop assessment criteria to evaluate their classmates' explanations or answers.
A13	0.58	Peer and self-assessment	In my classes I ask students to comment on their classmates' ideas or answers.
A14	0.35	Scaffolding and learning progression	When I assess students, I consider different levels of complexity to allow students progress in their understanding about a core idea.

Note: The items were sorted based on the theoretical subcategory used to theorize assessment literacy.

Table 23

Detail of the Items Included in the Third Factor Related to Teachers' Assessment Practices in Inquiry of the QALMBT-Generic for the Data Set Without Outliers

Item	Factor Loading	Theoretical dimension (Knowledge of…)	Statement
A20	0.65	Disciplinary knowledge and PCK	In order to assess students, I include laboratory activities that reinforce students' understanding of the core ideas addressed in class.
A6	0.46	Disciplinary knowledge and PCK	I use assessments to measure how students carry out investigations.
A1	0.34	Disciplinary knowledge and PCK	When I assess students learning, I evaluate whether students understand that knowledge may change in light of new evidence.
A5	0.32	Grading	I explain to students the criteria that I will use to evaluate their understanding.
A11	0.32	Peer and self-assessment	I ask students to test their hypothesis by identifying relationships between variables.

Note: The items were sorted based on the theoretical subcategory used to theorize assessment literacy.

Regarding the QALMBT-Modeling, which measured ISTs' assessment literacy specifically when teaching science with models, three factors were retained from the EFA conducted for the QALMBT-Modeling. Two of the three factors were similar to the factors identified in the

QALMBT-Generic for the data set without outliers, and they followed a similar trend; however, the third factor included items that focused on the interpretation of assessment and assessment criteria to help students reshape their ideas when thinking with models. It must also be pointed out that the factors identified for the QALMBT-Modeling with the data set with the outliers showed the same number of factors that the data set without outliers and included similar items in each of them.

The first factor included nineteen items, referred to as *Implementation of strategies to promote the elicitation and assessment of students' models.* Each of the theoretical dimensions suggested by the literature were included in this factor, and the majority of them included at least two items. In other words, this factor referred to a wide range of practices that ISTs used i) to facilitate the expression of a model, ii) to design formative and summative assessment instruments, iii) to prompt students to construct their understanding of a model, and iv) to facilitate the enrichment and progression of students' learning with a model. Table 24 shows the items that constitute this factor. The second factor included seven items and was conceptualized as *Self and peer assessment of generated models* (Table 25). This factor was related to engaging students in activities that promote the evaluation of their own models and those of their peers, for example, when judging the quality of elicited models in the classroom. The majority of the items were related to the theoretical dimension of engaging students in their own assessment of their models, as is indicated in Table 25.

Table 24

Detail of the Items Included in the First Factor Related to the Implementation of Strategies to Promote the Elicitation and Assessment of Students' Models of the QALMBT-Modeling

Item	Factor Loading	Theoretical dimension (Knowledge of…)	Statement
B26	0.68	Disciplinary knowledge and PCK	I use assessments to measure students' understanding of the models generated by them.
B20	0.34	Disciplinary knowledge and PCK	In order to assess students, I include laboratory activities that require the construction of models by students.
B31	0.65	Assessment purpose, content and methods	I use assessment to evaluate the internal consistency or coherence of various models constructed by a student.
B15	0.35	Assessment purpose, content and methods	I am able to translate the curriculum goals into clear specific tasks to evaluate students' models.
B28	0.74	Grading	When I develop distractors (incorrect or inferior alternatives) in a test, I have in mind the different alternatives or inaccurate ideas that students might have for the model in question.
B32	0.65	Grading	When I develop assessment instruments to assess students' models, I think beforehand how I will interpret the results
B19	0.42	Feedback	After a summative assessment, I clarify common students' misconceptions or alternative ideas about their generated models in class.
B16	0.38	Feedback	I communicate the results of the assessment in order to help each student achieve a better understanding of the expected model that I want them to learn.
B33	0.73	Feedback	I use assessment to give formative feedback to students about the phenomenon that they modeled.
B21	0.65	Assessment interpretation and communication	I use assessment to judge students' understanding about the phenomenon to be modeled.
B24	0.34	Assessment interpretation and communication	I use the results generated from formative assessment to adjust the content of my lessons regarding the model that I expect student to learn.
B34	0.79	Assessment interpretation and communication	I use assessment to locate evidence about the missing elements in a model that students have not understood.
B29	0.65	Peer and self-assessment	I challenge my students to show evidence to support their claims about their models.
B30	0.69	Peer and self-assessment	I use the results of the assessment to coach a student when she/he/they are having problems understanding a model.
B22	0.55	Assessment ethics	When students express their claims in front of the classroom, I establish classroom norms to promote a safe expression of students' ideas about their models.

Table 24 (continued)

Item	Factor Loading	Theoretical dimension (Knowledge of…)	Statement
B35	0.46	Assessment ethics	I tailor assessment in order to give all students the best opportunities to express their understanding about the model under study.
B25	0.68	Scaffolding and learning progression	I encourage students to use their pre-existing ideas in order to help them to construct an initial model than can be enriched later.
B9	0.33	Scaffolding and learning progression	I design scaffolded assignments or tasks that progress in complexity in order to assess students' understanding about the model under study.
B12	0.31	Scaffolding and learning progression	When I make an attempt for students to understand a model, I organize the content in my lessons following a sequence which considers how student understanding can evolve over a span of time.

Note: The items were sorted based on the theoretical subcategory used to theorize assessment literacy

The third factor included items related to the criteria of guiding students and providing information about the curricular model, scientific core ideas, or performance that students must achieve (Table 26). The name assigned to this factor was *Communication of assessment criteria to assess students' models* and was comprised of items related to the teacher explaining the criteria used to evaluate students' models, informing students in advance about the criteria that will be used to assess their models, and using assessment to compare students' enrichment and refinement of their models. It is worth remembering that each of the items included in each factor are based on the EFA and their conceptualization is based on an overall interpretation. Therefore, I acknowledge that the interpretation of individual items must be carefully judged.

Table 25

Detail of the Items Included in the Second Factor Related to Self and Peer Assessment of Generated Models of the QALMBT-Modeling

Item	Factor Loading	Theoretical dimension (Knowledge of…)	Statement
B1	0.38	Disciplinary knowledge and PCK	When I assess students learning, I evaluate whether students understand that models can be refined based on new evidence.
B4	0.76	Peer and self-assessment	I challenge my students to develop assessment criteria to evaluate the models constructed by their classmates.
B17	0.61	Peer and self-assessment	I teach students how to judge the quality of their explanations based on the consistency of their models.
B13	0.52	Peer and self-assessment	In my classes I ask students to comment on the models created by their classmates.
B11	0.47	Peer and self-assessment	I ask students to test their hypothesis by using their model.
B10	0.30	Assessment purpose, content and methods	I develop different kinds of assessment to evaluate students' reasoning by using models.
B14	0.39	Scaffolding and learning progression	When I assess students, I allow them to refine their models in order to help them reach different levels of complexity about the phenomenon that they are modeling.

Note: The items were sorted based on the theoretical subcategory used to theorize assessment literacy.

Table 26

Detail of the Items Included in the Third Factor Related to Communication of Assessment Criteria to Assess Student's Models of the QALMBT-Modeling

Item	Factor Loading	Theoretical dimension (Knowledge of…)	Statement
C5	0.74	Grading	I explain to students the criteria that I will use to evaluate their models.
C7	0.34	Assessment interpretation and communication	I use results from an assessment to compare how students' ideas about a model have been reshaped.
C8	0.80	Assessment ethics	When I develop summative assessment, I inform students in advance about the criteria that I will use to assess their models.

Note: The items were sorted based on the theoretical subcategory used to theorize assessment literacy.

Based on the EFA, it was identified that the QALMBT-Generic only included one factor; however, after removing outliers, this instrument seems to include three dimensions. Three

factors were also identified in the QALMBT-Modeling, where the third factor was related to informing assessment criteria. The third factor for QALMBT-Generic; however, was related to teacher assessment practices in inquiry. The identification of factors, particularly from the QALMBT-Modeling, is of particular interest to identify the components that might be related to assessment literacy in MBT among ISTs. In other words, the factors related to i) the implementation of strategies to promote the elicitation and assessment of students' models, ii) self and peer assessment of generated models, and iii) communication of assessment criteria to assess students' models are informative in terms of a baseline of a variety of assessment practices that teachers implement in their pedagogy when engaging students in modeling. Moreover, these groups of assessment strategies are also useful to answer research question 2 and identify if some of these group of practices changed after attending the OPDC.

In relation to the QALMBT- Epistemic, which measured ISTs' knowledge of the nature of models and modeling, three factors were identified. These factors followed the same trend for the data set with and without outliers. The three factors were assigned the names of i) *Generative tools for testing scientific knowledge,* ii) *Tentative nature of models*, and iii) *Multiplicity of scientific models.* The first factor refers to ISTs' knowledge of the purpose of models as generative tools that are developed and used to test hypotheses and assumptions, formulate new ideas, theories, and explanations about events, phenomena, or objects. This factor included 9 items (see Table 27) that were related to testing models, purpose of models and nature of models.

Table 27

Detail of the Items Included in the First Factor Related to Generative Tools for Testing Scientific Knowledge of the QALMBT-Epistemic

Item	Factor Loading	Theoretical dimension (Knowledge of…)	Statement
C12	0.63	Testing models	Models are developed to allow us to raise new questions and create new problems.
C2	0.42	Testing models	Hypotheses can be tested by using a model.
C10	0.34	Testing models	A model can be tested conceptually or non-experimentally.
C8	0.62	Purpose of models	Models are used to help formulate ideas and theories about scientific events, phenomena or objects.
C15	0.55	Purpose of models	Models can serve to transfer findings about a specific idea to another phenomenon.
C16	0.52	Purpose of models	Models can be used to test assumptions about something.
C9	0.37	Changing models	A model can be adjusted to reflect new findings.
C13	0.40	Nature of models	A model can be a mental image about a phenomenon that represents some entities of the original object under study.
C18	0.52	Nature of models	A model is a research tool that can be used to generate information.

The second factor reflected ISTs' knowledge of models as tools that result from a theoretical construction of the reality. In this sense as a theoretical construction, models can be refined after an iterative process of revision against empirical data in which scientists make testable predictions and check the models' explanatory power. The details of the five items for the second factor of QALMBT-Epistemic are included in Table 28. Finally, the third factor included items related to ISTs' knowledge of the diversity of models and the understanding that different models can coexist or represent a specific part of a phenomenon and therefore, models have scope and limitations (see Table 29).

Table 28

Detail of the Items Included in the Second Factor Related to Tentative Nature of Models of the QALMBT-Epistemic

Item	Factor Loading	Theoretical dimension (Knowledge of…)	Statement
C1	0.42	Nature of models	A model is a theoretical construction of reality.
C3	0.65	Changing models	Models need to be refined based on an iterative process in which empirical data compels the revision of the model.
C4	0.58	Purpose of models	Models can be designed with a main purpose of making testable predictions between variables.
C5	0.50	Changing models	Models could be changed when deduced hypotheses do not explain the original event, phenomena or object.
C11	0.38	Changing models	A characteristic of models is that they can be disproved when problems with its explanatory adequacy are identified by scientists.

Table 29

Detail of the Items Included in the Third Factor Related to Multiplicity of Scientific Models of the QALMBT-Epistemic

Item	Factor Loading	Theoretical dimension (Knowledge of…)	Statement
C6	0.48	Nature of models	A model differs in some degree from the reality.
C17	0.45	Testing models	Models need to be assessed to test their validity and fit with reality.
C19	0.48	Multiple models	A model can represent a specific part of a phenomenon under study rather than representing the entire phenomenon that constitutes in the real world.
C20	0.57	Multiple models	Each model has limitations making it necessary to generate several models to represent the reality.

These three factors identified in the QALMBT-Epistemic are aligned with Upmeir zu Belzen and Krüger (2010), Grünkorn, zu Belzen, and Krüger's (2014) and Krell and Krüger's (2016) framework related to models and modeling. As it was already mentioned in the elaboration of the QALMBT-Epistemic, this framework suggests that epistemological knowledge of models and modeling includes five aspects: the nature of models (e.g., models represent a target), multiple

models (e. g., multiple models can be generated to represent the same target), the purpose of models (e. g., models can be generated to describe, explain and predict), testing models (e. g., models are useful to test hypotheses), and changing models (e.g., models can be changed based on new evidence). The results from the EFA for the QALMBT-Epistemic only suggested three of the five sub-dimensions (Generative tools for testing scientific knowledge, ii) Tentative nature of models, and iii) Multiplicity of scientific models). A possible explanation to the identification of only three factors might be because the nature and purpose of models represent two aspects of the epistemological knowledge of models that are included in the majority of the dimensions. In this sense, the QALMBT-Epistemic suggests that the initial 5 sub-dimensions can be grouped in three broader categories.

4.2 Descriptive Statistics for the QALMT-Generic and QALMBT-Modeling Based on the Theoretical Dimensions Used to Define ALMBT

In this next section, I present the descriptive statistics for each section of the QALMBT questionnaire. By examining the overall trends for each theoretical dimension used to define ALMBT, I further characterize IST's assessment literacy in MBT (ALMBT) based on the results from the Chilean sample. The reader will recall that for each scale of the QALMBT questionnaire, a higher scale score meant a more frequent implementation of a type of assessment strategy in the classroom based on ISTs' self-report. The scale included five options: never (1), very rare (2), sometimes (3), frequently (4), and very frequently (5). The following section includes a table with the results for the QALMBT for each dimension used to determine assessment literacy. A comparison between the two countries is presented and the results are presented in a table for each theoretical dimension which details the percentage of agreement for

each item. I tis worth mentioning that for each item, the descriptive statistics for the QALMBT-Generic are presented first, followed by the descriptive statistics for the QALMBT-Modeling.

Disciplinary knowledge and PCK: As mentioned previously, disciplinary knowledge and PCK refers to teachers' knowledge of models and knowledge of how to use models and modeling in science education. Table 30 shows the descriptive statistics for the theoretical dimension of disciplinary knowledge and PCK for the Canadian and Chilean sample based on the results from the QALMBT-Generic and QALMBT-Modeling.

Table 30

Descriptive Statistics for the Items Included in the Dimension of Disciplinary Knowledge/ PCK

Statement		1	2	3	4	5	*M*	*SD*
1. When I assess students learning, I evaluate whether students understand that knowledge may change in light of new evidence.	Chile	1.8	6.7	25.65	36.0	29.8	3.9	1.0
	Canada	2.3	20.9	37.21	27.9	11.6	3.3	1.0
When I assess students learning, I evaluate whether students understand that models can be refined based on new evidence.	Chile	1.3	7.3	25.13	37.8	28.5	3.8	1.0
	Canada	4.7	18.6	32.56	30.2	14.0	3.3	1.1
6. I use assessments to measure how students carry out investigations.	Chile	3.6	10.9	25.13	29.8	30.6	3.7	1.1
	Canada	4.7	4.7	23.26	44.2	23.3	3.8	1.0
I use assessments to measure how students develop models to guide their investigations.	Chile	6.2	13.7	27.98	28.0	24.1	3.5	1.2
	Canada	7.0	16.3	48.84	27.9	0	3.0	.9
20. In order to assess students, I include laboratory activities that reinforce students' understanding of the core ideas addressed in class.	Chile	4.4	8.3	17.62	28.8	40.9	3.9	1.1
	Canada	0	0	25.58	34.9	39.5	4.1	.8
In order to assess students, I include laboratory activities that require the construction of models by students.	Chile	7.0	17.1	26.17	30.6	19.2	3.4	1.2
	Canada	4.7	20.9	44.19	18.6	11.6	3.1	1.0
26. I use assessments to measure students' understanding of core ideas.	Chile	.5	1.3	9.33	29.8	59.1	4.5	.8
	Canada	0	2.3	6.98	46.5	44.2	4.3	.7
I use assessments to measure students' understanding of the models generated by them.	Chile	3.1	8.0	23.32	36.5	29.0	3.8	1.0
	Canada	2.3	18.6	25.58	32.6	20.9	3.5	1.1

Note: The first row indicates the scale included in the questionnaire in which 1 = Never; 2 = Very rarely; 3 = Sometimes; 4 = Frequently; 5 = Very frequently. *M* = Mean; *SD* = Standard deviation.

Among disciplinary knowledge and PCK items, in both countries, ISTs scored higher for the QALMB-Generic than the QALMBT-Modeling. In terms of mean values for this dimension, it was found that Canadian ISTs selected the option "sometimes" more often for all items. This option means, for example, ISTs sometimes assess their students by challenging students to analyze new evidence (item 1) or to measure how students carry out investigations (item 6). In the case of Chile, ISTs stated they do these actions more frequently (values close to 4). Interestingly, almost half of the Canadian ISTs for item 6 "I use assessments to measure how students develop models to guide their investigations," answered that they do this action sometimes, and 27.9% of them indicated that do this action frequently when they teach with models. From Table 30 it seems that teachers in both countries more frequently use assessment to measure students' understanding of core ideas (item 26), but this frequency is lower when they are asked about models. In the case of the use of a laboratory (item 20), a large proportion of ISTs in Chile and Canada indicated they "frequently" and "very frequently" include laboratories to reinforce students understanding of core ideas, but the frequency drops when they are asked about how often they use assessment in laboratory activities that require the construction of models. In fact, more than 20% of the ISTs in both countries indicated they never or very rarely include this practice in their pedagogy.

Knowledge of assessment purpose, content, and methods: This dimension referred to teachers' knowledge of the role of assessment to judge students' performance or progress regarding the national/provincial curriculum. Table 31 summarizes the descriptive statistics for this theoretical dimension. Items related to the purpose of assessment criteria showed that the majority of teachers in both countries frequently use the provincial or national curriculum to

guide and design assessments (item 2). In other words, the results of the QALMBT-Generic and QALMBT-Modeling, respectively, showed that teachers align the scientific ideas or the expected models that students must learn based on the curriculum.

Table 31

Descriptive Statistics for the Items Included in the Dimension of Knowledge of Assessment Purpose, Content, and Methods

Statement		1	2	3	4	5	*M*	*SD*
2. I align my assessment with the goals of the provincial science curriculum when assessing students' ideas in science.	Chile	2.6	5.7	21.8	35.2	34.7	3.9	1.0
	Canada	2.3	4.7	9.3	44.2	39.5	4.1	.9
I align my assessment with the goals of the provincial science curriculum when assessing the expected models that students should learn.	Chile	2.9	3.4	18.4	43.5	31.9	4.0	.9
	Canada	0	4.7	27.9	39.5	27.9	3.9	.9
10. I develop different kinds of assessment to evaluate students' reasoning.	Chile	0.5	5.4	11.1	34.5	48.5	4.2	.9
	Canada	0	0	39.5	34.9	25.6	3.9	.8
I develop different kinds of assessment to evaluate students' reasoning by using models.	Chile	2.1	11.1	22.3	35.2	39.3	3.8	1.1
	Canada	9.3	7.0	46.51	32.6	4.7	3.2	1.0
15. I am able to translate the curriculum goals into clear specific tasks to guide my assessment activities.	Chile	1.0	4.4	16.6	34.5	43.5	4.2	.9
	Canada	0	2.3	14.0	46.5	37.2	4.2	.8
I am able to translate the curriculum goals into clear specific tasks to evaluate students' models.	Chile	4.4	10.4	25.4	35.0	24.9	3.7	1.1
	Canada	4.7	7.0	37.2	37.2	14.0	3.5	1.0
27. I assess how students make judgements in science based on reasoning.	Chile	1.3	4.7	20.7	32.1	41.2	4.1	1.0
	Canada	0	2.3	23.4	51.2	23.3	4.0	.8
I assess how students make judgements in science based on reasoning with a model.	Chile	1.3	11.9	29.5	35.5	21.8	3.6	1.0
	Canada	7.0	23.3	23.2	30.2	16.3	3.3	1.2
31. I use assessment to evaluate the internal consistency or coherence of students' ideas.	Chile	0.8	5.4	15.3	41.2	37.3	4.1	.9
	Canada	2.3	7.0	27.9	32.6	30.2	3.8	1.0
I use assessment to evaluate the internal consistency or coherence of various models constructed by a student.	Chile	4.2	12.4	29.8	34.5	19.2	3.5	1.1
	Canada	9.3	20.9	37.2	18.6	14.0	3.1	1.2

Note: The first row indicates the scale included in the questionnaire in which 1 = Never; 2 = Very rarely; 3 = Sometimes; 4 = Frequently; 5 = Very frequently. *M* = Mean; *SD* = Standard deviation.

For item 2, the mean values were close in both countries. For the QALMBT-Generic the means were 4.1 and 3.9 for Canada and Chile, whereas for the QALMBT-Modeling the means were 3.9 and 4.0, respectively. For the remaining items in this criterion, the mean values were lower when ISTs were asked about an MBT approach. For example, for item 15 related to curriculum goals, the number of ISTs in both countries who never or very rarely were able to translate the curriculum goals into clear specific tasks almost tripled in size in the case of the QALMBT-Modeling. In fact, the mean values for the QALMBT-Generic and the QALMBT-Modeling for Canada and Chile were respectively 3.7, 4.2, and 3.5, 4.2. Regarding the development of assessment instruments, in item 10, almost 40% of the Canadian ISTs answered that they sometimes develop different instruments to evaluate students' reasoning (39.5%) or to evaluate students' reasoning using models (46.5%). More than 15% of the participants indicated that they never or very rarely develop a variety of assessments to assess students' models. The Chilean sample showed a higher percentage of teachers who frequently or very frequently implemented this indicator (item 10) in their general approach to teaching science (more than 80%), and almost 75% of respondents indicated they do evaluate students' reasoning when they teach with models. The lowest mean values for the QALMBT-Modeling for the dimension purpose of assessment were observed for question 31. Surprisingly, one out of three of the Canadian ISTs indicated they never or very rarely use assessment to evaluate students' internal consistency or coherence of various models. In the case of Chile, one out of six ISTs selected that they never or very rarely evaluate students' internal consistency in their models, whereas more than 55% of them indicated they do it frequently or very frequently. This result contrasts with the Canadian sample in which almost one out of three ISTs selected the option frequently or very frequently.

Knowledge of grading: This theoretical dimension referred to teachers' capacity to construct and implement scoring techniques and different instruments to assess their students. In this dimension, teachers' self-report on this dimension were much higher for the generic approach in both countries. Table 32 shows the descriptive statistics for the theoretical dimension of knowledge of grading.

Table 32

Descriptive Statistics for the Items Included in the Dimension of Knowledge of Grading

Statement		1	2	3	4	5	*M*	*SD*
5. I explain to students the criteria that I will use to evaluate their understanding.	Chile	0.3	1.0	7.0	24.6	67.1	4.6	0.7
	Canada	0	0	7.0	55.8	37.2	4.3	0.6
I explain to students the criteria that I will use to evaluate their models.	Chile	2.9	3.4	15.3	28.5	50.0	4.2	1.0
	Canada	2.3	9.3	14.0	44.2	30.2	3.9	1.0
23. I design different scoring tools (e.g., rubrics, checklists, standards) to judge students work.	Chile	3.4	3.4	12.2	24.4	56.7	4.3	1.0
	Canada	0	11.6	11.4	39.5	37.2	4.0	1.0
I design different scoring tools (e.g., rubrics, checklists, standards) to evaluate models generated by students.	Chile	2.3	8.3	18.1	17.2	44.0	4.0	1.1
	Canada	4.7	16.3	23.3	25.6	30.2	3.6	1.2
28. When I develop distractors (incorrect or inferior alternatives) in a test, I have in mind the different alternatives or inaccurate ideas that students might have for the content that they are studying.	Chile	1.8	3.1	11.4	30.3	53.4	4.3	0.9
	Canada	9.3	18.6	14.0	32.6	25.6	3.5	1.3
When I develop distractors (incorrect or inferior alternatives) in a test, I have in mind the different alternatives or inaccurate ideas that students might have for the model in question.	Chile	2.9	8.8	25.0	31.1	32.4	3.81	1.1
	Canada	16.3	23.3	25.6	18.6	16.3	2.95	1.3
32. When I develop assessment instruments to assess students' understanding, I think beforehand how I will interpret the results.	Chile	3.4	4.9	15.3	31.9	44.6	4.1	1.0
	Canada	2.3	2.3	23.3	37.2	34.9	4.0	1.0
When I develop assessment instruments to assess students' models, I think beforehand how I will interpret the results	Chile	5.7	7.8	23.1	34.5	29.0	3.7	1.1
	Canada	7.0	9.3	41.9	20.9	20.9	3.4	1.1

Note: The first row indicates the scale included in the questionnaire in which 1 = Never; 2 = Very rarely; 3 = Sometimes; 4 = Frequently; 5 = Very frequently. *M* = Mean; *SD* = Standard deviation.

For item 5, for example, the vast majority of ISTs in both countries stated that they frequently explain the criteria used to evaluate the understanding and models to their students. Similar results were reported in item 23 regarding the variety of scoring tools that ISTs utilize to judge students' work and models. Regarding the construction of distractors in a test and ISTs' awareness of students' alternatives or inaccurate ideas (item 28), almost a third of the Canadian ISTs selected that they never or very rarely reflect on these ideas, and a quarter of them sometimes do it. In the case of the Chilean ISTs, their self-report for this item is considerably higher, which is also reflected in the mean values included in Table 32 for the QALMBT-Generic and QALMBT-Modeling, 4.3, and 3.8, respectively, whereas for the Canadian sample, the mean values were 3.46 and 2.95. Finally, for the item related to whether ISTs they think will grade the students' answers before giving the assessment (item 32), more than 40% of the Canadian ISTs indicated they frequently or very frequently think about the interpretation of the results obtained from assessment when the activities involve models. For the Chilean sample, almost 65% of the respondents chose the same option.

Knowledge of feedback: This theoretical dimension emphasized the utility of assessment as a tool to communicate feedback to students. Participants in both countries showed they frequently communicated feedback to their students when they teach generally. The mean values were close to or over 4 for all the items related to this dimension. Nevertheless, specifically for questions 16 and 33, Canadian ISTs scored lower than Chilean ISTs with mean values of 3.6 and 3.5, respectively. The percentage of ISTs' answers for each item is detailed in Table 33.

Table 33

Descriptive Statistics for the Items Included in the Dimension of Knowledge of Feedback

Statement		1	2	3	4	5	*M*	*SD*
3. For those areas that students have difficulty in comprehending, I promote the generation of a consensus explanation that helps students have a similar understanding of the core ideas.	Chile	0.5	3.1	15.0	38.9	42.3	4.2	0.8
	Canada	2.3	4.7	25.6	37.2	30.2	3.9	1.0
For those areas that students have difficulty in comprehending, I promote the generation of a consensus model that helps students have a similar understanding of the phenomenon under study	Chile	1.3	5.2	16.8	47.2	29.5	4.0	0.9
	Canada	2.3	7.0	20.9	46.5	23.3	3.8	1.0
16. I communicate the results of the assessment in order to help each student refine their initial ideas.	Chile	0.8	2.6	9.8	32.9	53.9	4.4	0.8
	Canada	0	0	23.3	46.5	30.2	4.1	0.7
I communicate the results of the assessment in order to help each student achieve a better understanding of the expected model that I want them to learn.	Chile	1.3	5.7	13.2	33.9	45.9	4.2	1.0
	Canada	2.3	16.3	20.9	39.5	20.9	3.6	1.1
19. After a summative assessment, I clarify students' wrong answers.	Chile	1.0	3.4	9.3	20.7	65.5	4.5	0.9
	Canada	0	14.0	16.3	25.6	44.2	4.0	1.1
After a summative assessment, I clarify common students' misconceptions or alternative ideas about their generated models in class.	Chile	1.6	5.2	10.4	29.0	54.9	4.3	1.0
	Canada	4.7	9.3	18.6	27.9	39.5	3.9	1.2
33. I use assessment to give formative feedback to students about their understanding of the core ideas studied in class.	Chile	1.0	2.1	9.6	26.4	60.9	4.4	0.8
	Canada	0	4.7	9.3	46.5	39.5	4.2	0.8
I use assessment to give formative feedback to students about the phenomenon that they modeled.	Chile	3.1	7.8	21.8	36.0	31.4	3.8	1.0
	Canada	4.7	14.0	30.2	32.6	18.6	3.5	1.1

Note: The first row indicates the scale included in the questionnaire in which 1 = Never; 2 = Very rarely; 3 = Sometimes; 4 = Frequently; 5 = Very frequently. *M* = Mean; *SD* = Standard deviation.

In the case of question item 16 related to communicating feedback to help students achieve a better understanding of the expected model, the percentage of Canadian teachers who selected that they very rarely communicate the results of assessment increased from 0 to 16.3 when the question was related to models. In the case of the Chilean sample, this percentage fluctuated

from 2.6 to 5.7 for the QALMBT-Generic and QALMBT-Modeling, respectively. Similarly, in question 33, concerning formative assessment, almost 20% of the Canadian ISTs selected that they never or very rarely used formative assessment to give feedback to their students. Moreover, more than 30% of the Canadian ISTs answered they do this action sometimes when they teach with models. This percentage was considerably higher when they answered the same question but from a generic approach to teaching: more than 85% of the ISTs in both countries selected the option frequently or very frequently communicate feedback. When teachers were asked about summative assessment (item 19), the proportions were similar for each country independent of the approach, but the mean for the QALMBT-Modeling was lower.

Knowledge of assessment interpretation and communication: This theoretical dimension referred to the interpretation of the results obtained from assessments to judge students' reasoning in science. The results show that Canadian ISTs scored particularly low as compared with the Chilean sample. When Canadian teachers were asked about MBT, the score in their answers dropped considerably in comparison to a general approach to teaching science. This result contrasts with Chilean teachers whose mean was close to 4 for each item. The item that showed the lowest score for the Canadian sample was item 7, which suggests that almost half of the teachers only once every month or once every unit compare how students' understanding of models has been reshaped (M_{Chile} = 3.7; M_{Canada} = 3.0). Answers in question 21 also fluctuated substantially. For example, more than 80% of the Chilean ISTs (M = 4.2) answered that they frequently or very frequently use assessment to judge students' understanding about the phenomenon, whereas slightly more than 55% of them indicated they do this action when their pedagogy involves models (M = 3.6). Regarding the Canadian sample, these percentages

fluctuated from more than 80% to 44%, with a mean value of 4.2 and 3.3 for the QALMBT-Generic and QALMBT-Modeling, respectively. In item 24, when teachers were asked about the use of formative assessment to adjust the content of their lessons, the mean values in Chile remained close to 4 for both, the QALMBT-Generic and QALMBT-Modeling, but in the case of Canada, the mean fluctuated from 4.2 for a general approach to teaching and only 3.4 for a MBT approach. Table 34 shows the results for for this theoretical dimension.

Table 34

Descriptive Statistics for the Items Included in the Dimension of Knowledge of Assessment Interpretation and Communication

Statement		1	2	3	4	5	*M*	*SD*
7. I use results from an assessment to compare how students' understanding about the topic under study has been reshaped.	Chile	1.3	6.2	20.0	40.7	31.9	4.0	0.9
	Canada	2.3	9.30	32.6	34.9	20.9	3.6	1.0
I use results from an assessment to compare how students' ideas about a model have been reshaped.	Chile	3.4	10.1	26.9	33.4	26.2	3.7	1.1
	Canada	9.3	16.3	46.5	23.3	4.7	3.0	1.0
21. I use assessment to judge students' understanding about the phenomenon under study.	Chile	1.8	1.8	13.7	36.3	46.4	4.2	0.9
	Canada	0	2.3	14.0	16.5	37.2	4.2	0.8
I use assessment to judge students' understanding about the phenomenon to be modeled.	Chile	3.9	10.9	28.8	35.8	20.7	3.6	1.1
	Canada	7.0	9.3	39.5	30.2	14.0	3.3	1.1
24. I use the results generated from formative assessment to adjust the content of my lessons regarding the core ideas under study.	Chile	1.8	3.1	12.7	32.6	49.7	4.3	0.9
	Canada	0	4.7	14.0	37.2	44.2	4.2	0.9
I use the results generated from formative assessment to adjust the content of my lessons regarding the model that I expect student to learn.	Chile	2.1	8.0	19.2	36.3	34.5	3.9	1.0
	Canada	4.6	14.0	32.6	32.6	16.3	3.4	1.1
34. I use assessment to locate evidence about the missing elements that students have not understood regarding the core ideas under study.	Chile	1.3	1.6	9.3	31.9	56.0	4.4	0.8
	Canada	0	2.3	30.2	32.6	34.9	4.0	0.9
I use assessment to locate evidence about the missing elements in a model that students have not understood.	Chile	2.9	6.2	23.1	35.0	32.9	3.9	1.0
	Canada	7.0	11.6	41.9	23.3	16.3	3.3	1.1

Note: The first row indicates the scale included in the questionnaire in which 1 = Never; 2 = Very rarely; 3 = Sometimes; 4 = Frequently; 5 = Very frequently. *M* = Mean; *SD* = Standard deviation.

Knowledge of Peer and Self-Assessment: This dimension referred to teachers' capacity to engage students in assessment, for example, by evaluating their very own explanations and models. Table 35 summarizes the descriptive statistics for each of the items included in this theoretical dimension.

Table 35

Descriptive Statistics for the Items Included in the Dimension of Knowledge of Peer and Self-Assessment

Statement		1	2	3	4	5	*M*	*SD*
4. I challenge my students to develop assessment criteria to evaluate their classmates' explanations or answers.	Chile	10.4	25.4	21.2	29.3	13.7	3.1	1.2
	Canada	18.6	34.9	20.9	20.9	4.7	2.6	1.2
I challenge my students to develop assessment criteria to evaluate the models constructed by their classmates.	Chile	15.3	21.8	26.4	26.7	9.8	2.9	1.2
	Canada	25.6	25.6	23.3	18.6	7.0	2.6	1.3
11. I ask students to test their hypothesis by identifying relationships between variables.	Chile	1.8	4.9	20.5	38.3	34.5	4.0	1.0
	Canada	0	11.6	27.9	37.2	23.3	3.7	1.0
I ask students to test their hypothesis by using their model.	Chile	3.6	11.9	23.3	39.6	21.6	3.6	1.1
	Canada	11.6	23.3	37.2	18.6	9.3	2.9	1.1
13. In my classes I ask students to comment on their classmates' ideas or answers.	Chile	2.3	9.3	17.4	32.9	38.1	4.0	1.1
	Canada	9.3	16.3	25.6	27.9	20.9	3.3	1.3
In my classes I ask students to comment on the models created by their classmates.	Chile	7.0	13.5	30.8	27.2	21.5	3.4	1.2
	Canada	9.3	34.9	23.3	20.9	11.6	2.9	1.2
17. I teach students how to judge the quality of their explanations based on the consistency of their ideas.	Chile	4.4	6.2	18.9	36.0	34.5	3.9	1.1
	Canada	2.3	11.6	23.3	39.5	23.3	3.7	1.0
I teach students how to judge the quality of their explanations based on the consistency of their models.	Chile	5.4	10.9	27.0	32.9	23.8	3.6	1.1
	Canada	4.7	23.3	32.6	27.9	11.6	3.2	1.1
29. I challenge my students to show evidence to support their claims about the content that they are studying.	Chile	2.3	3.8	13.9	31.3	48.7	4.2	1.0
	Canada	0	2.3	14.0	46.5	37.2	4.2	0.8
I challenge my students to show evidence to support their claims about their models.	Chile	4.4	8.8	22.5	35.2	29.0	3.8	1.1
	Canada	7.0	11.6	23.3	30.2	27.9	3.6	1.2

Note: The first row indicates the scale included in the questionnaire in which 1 = Never; 2 = Very rarely; 3 = Sometimes; 4 = Frequently; 5 = Very frequently. *M* = Mean; *SD* = Standard deviation.

Challenging students to develop assessment criteria (item 4) seems to be a practice that ISTs do not implement very often. In both countries, there was almost no difference between the QALMBT-Generic and QALMBT-Modeling on this dimension. The mean values in the QALMBT-Generic and QALMBT-Modeling were close to 3 for the Chilean sample and close to 2.5 in the case of the Canadian sample. When teachers were asked about encouraging students to test their own hypothesis by identifying relationships between variables, responses in item 11 revealed that more than 70% of ISTs in Chile indicated they frequently or very frequently engage students in this form of assessment in class and more than 70% of them also indicated that they did it when the activity involved models. Regarding Canadian science teachers, less than 30% of them answered that they implement this strategy frequently or very frequently when teaching with models, but more than 60% of them indicated that they use it in a general approach to teaching science. For items 13 and 17 related to asking students to comment on others' ideas or judge the quality of their own explanations, respectively, the mean values showed that science teachers in Chile do this action frequently during a general approach to teaching science, but the mean value decreased from almost 4 to close to 3.5 in MBT. This tendency was also observed with the Canadian ISTs whose mean remained close to 3 for the QALMBT-Generic and QALMBT-Modeling. Findings obtained from these four items (4, 11, 13 and 17) might suggest that teachers in both countries do not engage their students too often in the process of peer and self-assessment. It appears plausible that many teachers do not challenge their students to criticize and revise their own or others' ideas in class. Item 29 was the only item that had a mean of over 4 in both countries for the QALMBT-Generic; however, this value decreased at least .5 points in MBT. This result might indicate that teachers often ask their students to justify their claims with evidence, but teachers do not request their students to support their models.

Knowledge of Assessment Ethics. This theoretical dimension was related to teachers' capacity to provide each student equitable opportunities to participate in class and reach the same level of understanding as his/her peers. Table 36 displays the descriptive statistic for this dimension.

Table 36

Descriptive Statistics for the Items Included in the Dimension of Knowledge of Assessment Ethics

Statement		1	2	3	4	5	*M*	*SD*
8. When I develop summative assessment, I inform students in advance about the criteria that I will use to assess understanding in class.	Chile	0.5	2.1	8.6	23.1	65.8	4.5	0.8
	Canada	0	7.0	7.0	37.2	48.8	4.3	0.9
When I develop summative assessment, I inform students in advance about the criteria that I will use to assess their models.	Chile	3.4	5.4	14.5	25.9	50.8	4.2	1.0
	Canada	2.3	14.0	14.0	34.9	34.9	3.9	1.1
22. When students express their claims in front of the classroom, I establish classroom norms to promote a safe expression of students' ideas about the disciplinary core ideas in science.	Chile	0.8	0.3	7.3	22.3	69.4	4.6	0.7
	Canada	2.3	2.3	14.0	32.6	48.8	4.2	0.9
When students express their claims in front of the classroom, I establish classroom norms to promote a safe expression of students' ideas about their models.	Chile	1.3	5.4	16.6	28.2	48.5	4.2	1.0
	Canada	4.7	9.3	16.3	30.2	39.5	3.9	1.2
30. I use the results of the assessment to coach a student when she/he/they are having problems understanding a disciplinary core idea in science.	Chile	1.6	4.9	14.8	33.4	45.3	4.2	1.0
	Canada	0	2.3	23.3	32.6	41.9	4.1	0.9
I use the results of the assessment to coach a student when she/he/they are having problems understanding a model.	Chile	2.6	10.6	23.6	34.7	28.5	3.8	1.1
	Canada	4.7	14.0	25.6	23.3	32.6	3.7	1.2
35. I tailor assessment in order to give all students the best opportunities to express their understanding about the disciplinary core ideas under study.	Chile	5.2	12.2	20.5	25.5	36.8	3.8	1.2
	Canada	0	9.3	20.9	41.9	27.9	3.9	0.9
I tailor assessment in order to give all students the best opportunities to express their understanding about the model under study.	Chile	6.2	16.8	26.4	29.5	21.0	3.4	1.2
	Canada	7.0	16.3	34.9	25.6	16.3	3.3	1.1

Note: The first row indicates the scale included in the questionnaire in which 1 = Never; 2 = Very rarely; 3 = Sometimes; 4 = Frequently; 5 = Very frequently. *M* = Mean; *SD* = Standard deviation.

In general, answers belonging to this dimension showed high mean values for the QALMBT-Generic in both countries. For example, in question 8, a high proportion of the ISTs answered they frequently and very frequently inform their students in advance about the criteria they will use to assess students' understanding and models. Similar results were observed in question 22. This item was related to the implementation of classroom norms to facilitate students' expression of core ideas in science and their models. Even though in both versions of the questionnaire the mean values were high, the mean values for the QALMBT-Modeling were considerably lower than the QALMBT-Generic. Regarding using assessment to couch students' problems with understanding disciplinary ideas or problems with understanding a model (item 30), more than 70% of the Canadian and Chilean ISTs answered that they frequently or very frequently use the results of the assessment to monitor students' understanding in the classroom. In the case of the answers related to modeling, see table 36, the percentage of ISTs who indicated that they never or very rarely implement this assessment strategy doubled in the case of the Chilean sample and increased considerably in the case of the Canadian sample, which corresponded to 13.2% and 18.6%, respectively. Item 35 showed the lowest mean values in this category: Chilean ISTs showed a mean of 3.8 for the QALMBT-Generic and 3.4 for the QALMBT-Modeling, whereas in Canada the means were 3.88 for a generic approach to teaching science and 3.3 for MBT.

Knowledge of Scaffolding and Learning Progressions: Finally, this theoretical dimension referred to the assessment strategies that teachers used to help students progress in their learning. When teachers were asked about scaffolding and learning progressions, the mean values for the QALMBT-Modeling decreased considerably for several items in comparison to a general approach to teaching science. Table 37 details the descriptive statistics this dimension.

Table 37

Descriptive Statistics for Dimension of Knowledge of Scaffolding and Learning Progression

Statement		1	2	3	4	5	*M*	*SD*
9. I design scaffolded assignments or tasks that progress in complexity in order to assess students' understanding about disciplinary core ideas in science.	Chile	0.8	3.4	17.6	38.6	39.6	4.1	0.9
	Canada	0	7.0	11.6	48.8	32.6	4.1	0.9
I design scaffolded assignments or tasks that progress in complexity in order to assess students' understanding about the model under study.	Chile	2.1	5.7	28.0	38.3	25.9	3.8	1.0
	Canada	2.3	14.0	25.6	34.9	23.3	3.6	1.1
12. When I make an attempt for students to understand a core idea in science, I organize the content in my lessons following a sequence which considers how student understanding can evolve over a span of time.	Chile	1.3	1.3	11.7	29.5	56.2	4.4	0.8
	Canada	4.7	2.3	16.3	27.9	48.8	4.1	1.1
When I make an attempt for students to understand a model, I organize the content in my lessons following a sequence which considers how student understanding can evolve over a span of time.	Chile	1.3	5.7	16.3	38.6	38.1	4.1	0.9
	Canada	7.0	4.7	23.3	25.6	39.5	3.9	1.2
14. When I assess students, I consider different levels of complexity to allow students to progress in their understanding about a core idea.	Chile	0.5	2.9	13.2	32.9	50.5	4.3	0.8
	Canada	0	0	25.6	39.5	34.9	4.1	0.8
When I assess students, I allow them to refine their models in order to help them reach different levels of complexity about the phenomenon that they are modeling.	Chile	3.6	0.3	26.2	38.1	22.8	3.7	1.0
	Canada	9.3	14.0	32.6	25.6	18.6	3.3	1.2
18. I deconstruct a task or objective from the science curriculum into smaller instructional learning experiences that assess students' progression with a core idea.	Chile	0.5	3.1	11.1	34.7	50.5	4.3	0.8
	Canada	2.3	2.3	7.0	34.9	53.5	4.3	0.9
I deconstruct a task or objective from the science curriculum into smaller instructional learning experiences that assess students' progression with their models.	Chile	3.1	11.1	24.9	36.8	24.1	3.7	1.1
	Canada	4.7	14.0	39.5	25.6	16.3	3.3	1.1
25. I encourage students to use their pre-existing ideas in order to help them to construct an initial explanation that can be enriched later.	Chile	1.0	2.9	9.3	28.2	58.6	4.4	0.9
	Canada	2.3	2.3	25.6	32.6	37.2	4.0	1.0
I encourage students to use their pre-existing ideas in order to help them to construct an initial model than can be enriched later.	Chile	2.1	6.7	18.4	34.5	38.3	4.0	1.0
	Canada	4.7	20.9	25.6	27.9	20.9	3.4	1.1

Note: The first row indicates the scale included in the questionnaire in which 1 = Never; 2 = Very rarely; 3 = Sometimes; 4 = Frequently; 5 = Very frequently. *M* = Mean; *SD* = Standard deviation.

For example, in question 14, Canadian ISTs showed a mean of 4.1 when they were asked about different levels of complexity to foster students progression in their individual understanding of a core idea. Similarly, for the same question, Chilean ISTs showed a mean value of 4.30. In the case of an MBT approach, the mean dropped to 3.3 and 3.7, respectively, for each country. Participants' answers in question 18 that asked how often teachers deconstruct a task or objective from the science curriculum into smaller instructional learning experiences to assess the progression of models, revealed a mean value of almost 0.7% lower in the case of Chilean ISTs, and 1 point lower for the Canadian sample in comparison to the results from the QALMBT-Generic. In another example, in relation to the incorporation of students' pre-existing ideas (item 25), in both countries, ISTs indicated they frequently include them to help students construct initial explanations; however, in the case of the Canadian ISTs, their answers varied considerably from a general approach ($M = 4.0$) to an MBT approach ($M = 3.4$).

In summary, the main findings of the QALMBT revealed that in general, Chilean ISTs had a higher baseline ALMBT as compared to Canadian ISTs. This finding was reflected in higher mean values for each dimension in the questionnaire. Moreover, for each of the dimensions and each of the items, the scores in a generic approach were higher in comparison to an MBT approach for both countries. Interestingly, the results from the dimension related to disciplinary knowledge and PCK showed that ISTs in both countries use assessment more often to assess students' understanding of core ideas in science (means over 4); however, when they were asked about engaging students in the generation and evaluation of models (e. g., items 1 and 6), their means were close to 3 (sometimes). Similarly, the results from the theoretical dimension of knowledge of the purpose of assessment the items related to i) developing different assessment

instruments to evaluate students' reasoning by using models (item 10), ii) assessing how students make judgments in science based on reasoning with models (item 27), and using assessment to evaluate the internal consistency of various models constructed by a student (item 31) reported values close to 3 for the Canadian sample and close to 3.5 for the Chilean sample even though in a generic approach to teaching science the means were close to 4. In relation to the dimension of knowledge of grading, overall, the majority of the items included showed high means with values higher than 4 for each item for the QALMBT-Generic and QALMBT-Modeling. Nevertheless, in the case of the Canadian sample, their results showed that for item 28 ("When I develop distractors in a test, I have in mind the different alternative ideas that students might have for the model in question") and 32 ("When I develop assessment instruments to assess students' models, I think beforehand how I will interpret the results.") the majority of ISTs sometimes used this strategy. In the case of the Chilean sample, even though the results from the QALMBT-Modeling were lower than the QALMBT-Generic, these results were considerably higher in comparison to the Canadian sample (see, for example, item 28).

The indicators included in the theoretical dimension of knowledge of feedback showed the highest means for the QALMBT-Generic and QALMBT-Modeling. Moreover, similar mean values were identified when teachers used a generic and MBT approach to teaching science, for both the Canadian and Chilean sample. The lowest mean value in this dimension was observed in the Canadian sample for item 33 in an MBT approach (3.5) related to using assessment to give formative feedback to students about the phenomenon that they modeled, whereas in a generic approach it was considerably higher (4.2). In the case of the dimension of knowledge of assessment interpretation and communication the items related to using assessment to compare

how students' ideas about a model have been reshaped (item 7) and using assessment to judge students' understanding about the phenomenon to be modeled showed the lowest values for this dimension for both countries with values that fluctuated from 3 to 3.5. As it was identified in all the dimensions, the items included in the dimension of knowledge of assessment ethics reported high means for a generic approach (means over 4) but when being asked about models the results were slightly lower. The item related to tailoring assessment in order to give all students the best opportunities to express their understanding about the model under study was particularly low in the Canadian sample with a mean of 3.3 whereas for the Chilean sample was 3.4. When being asked about the dimension of knowledge of scaffolding and learning progression, two items (12 and 14) showed mean values that fluctuated from 3.3 to 3.7 in an MBT approach for both countries. In other words, the majority of Canadian and Chilean ISTs only sometimes allow their students to refine their models to reach different levels of complexity about the phenomenon to be modeled and they sometimes deconstruct a task from the science curriculum into smaller instructional learning experiences that assess students' progression with their models.

Interestingly, the dimension related to engaging students in assessment was the dimension that showed the lowest score for many of its items in both countries and in both a generic and MBT approach. Specifically, when being asked about and MBT approach, ISTs scored lower values than in a generic approach of teaching. The item related to challenging students to develop assessment criteria to evaluate their classmates' explanations/models constructed by their classmates, was one of the least often assessment practices used by ISTs with means of 2.6 and 3.1 for the Canadian and the Chilean sample in a generic approach and 2.6 and 2.9, respectively for an MBT approach.

4.3 Descriptive Statistics for the QALMBT-Epistemic

In this section of the questionnaire, a higher score represented a more sophisticated knowledge about models and their function in science. Teachers who selected totally disagree, disagree, or undecide had a more basic or limited epistemological understanding, as the scale used was a 1-5 Likert scale with 1 being strongly disagree. Understanding the nature of modeling in science was important to this investigation to ascertain its relationship to practices in order to answer research question 1.

Nature of Models: This theoretical dimension referred to teachers' knowledge about the nature of models as theoretical reconstructions to represent a target (e. g., phenomenon, event, mechanism, etc.). Table 38 displays the descriptive statistics for the items related to the nature of models.

Table 38

Descriptive Statistics for QALMBT-Epistemic for Items Related to the Nature of Models

Statement		1	2	3	4	5	*M*	*SD*
1. A model is a theoretical construction of reality.	Chile	3.8	7.0	12.4	38.4	38.4	4.0	1.1
	Canada	2.3	0	0	53.5	44.2	4.4	0.7
6. A model differs in some degree from the reality.	Chile	5.4	13.4	15.6	32.8	32.8	3.7	1.2
	Canada	0	11.6	7.0	30.2	51.2	4.2	1.0
13. A model can be a mental image about a phenomenon that represents some entities of the original object under study.	Chile	2.0	4.6	15.1	39.2	39.2	4.1	1.0
	Canada	2.3	2.3	16.3	41.9	37.2	4.1	0.9
18. A model is a research tool that can be used to generate information	Chile	2.2	4.0	6.7	33.9	52.2	4.3	0.9
	Canada	2.3	4.7	11.6	32.6	48.8	4.2	1.0

Note: The first row indicates the scale included in the questionnaire in which 1 = Strongly disagree; 2 = Disagree; 3 = Undecided; 4 = Agree; 5 = Strongly disagree. *M* = Mean; *SD* = Standard deviation.

Overall, according to the questionnaire, QALMBT-Epistemic, most teachers in both countries understood that models can be a mental image about a phenomenon that represent some aspects of reality (item 13); however, more than 15% of Chilean teachers chose undecided on whether they agreed or disagreed with the statement; models can be a mental image, whereas only 7% of the Canadian sample were undecided. Interestingly, more than 10% of the Chilean ISTs scored that they strongly disagreed or disagreed with item 1 that a model is a theoretical construction of reality, whereas almost 12% of them were undecided. By contrast, more than 90% of the Canadian ISTs indicated they agreed or strongly agreed with the statement that a model is a theoretical construction of reality. For item 6 that a model differs to some degree from reality, Chilean ISTs showed a considerably lower mean value in comparison to the Canadian sample (*M* = 3.7 and 4.2, respectively). For example, almost 15.6% of the Chilean ISTs were undecided that a model differs to some degree from reality, and nearly 20% of them selected that they disagreed or strongly disagreed with the statement. In the case of Canada, nearly 7% of those surveyed was undecided, and 11.6% disagreed with this item. Regarding the use of models as a research tool (item 18), the vast majority of the ISTs in both countries agreed with the notion that models can be used to generate information. Only 7% of the Canadian and the Chilean ISTs did not conceptualize models as research tools.

Multiple Models: This dimension related to teachers' knowledge of the limitations and scope of models and the utility that multiple models for explaining, for example, the same target from different perspectives. Table 39 shows the results from the QALMBT-Epistemic for the items related to multiple models. The results of the items related to the multiplicity of models showed that ISTs in Chile and Canada had a rich knowledge of this dimension. For example, for each

item included in this dimension, the mean values were larger than 4, and more than 75% of the participants indicated they agreed or strongly agreed with every item.

Table 39

Descriptive Statistics for QALMBT-Epistemic for Items Related to Multiple Models

Statement		1	2	3	4	5	*M*	*SD*
7. Testing competing scientific models gives better insight into the explanatory scope of each of them.	Chile	1.1	1.9	14.0	40.0	44.1	4.2	0.8
	Canada	2.3	0	4.7	42.0	51.2	4.4	0.8
14. Competing models can coexist in science to represent the same object, phenomenon or system.	Chile	2.4	4.8	14.0	42.2	36.6	4.1	1.0
	Canada	0	4.7	16.3	23.3	55.8	4.3	0.9
19. A model can represent a specific part of a phenomenon under study rather than representing the entire phenomenon that constitutes in the real world.	Chile	1.1	2.4	11.6	36.3	48.7	4.3	0.8
	Canada	0	2.3	7.0	23.3	67.4	4.6	0.7
20. Each model has limitations making it necessary to generate several models to represent the reality.	Chile	1.6	4.0	16.1	33.8	45.4	4.2	0.9
	Canada	2.3	4.7	11.6	44.2	37.2	4.1	0.9

Note: The first row indicates the scale included in the questionnaire in which 1 = Strongly disagree; 2 = Disagree; 3 = Undecided; 4 = Agree; 5 = Strongly disagree. *M* = Mean; *SD* = Standard deviation.

For item 7, in which teachers were asked if testing competing scientific models gives a better insight into the explanatory power of each of them, both countries showed a high percentage of agreement. Agreement notwithstanding, Chilean ISTs showed three times more undecided responses than the Canadian sample (13.98% and 4.65%, respectively). When teachers were asked about competing models in item 14, almost 15% of the Chilean and Canadian ITSs still remained undecided that different models can coexist to represent the same target. A similar percentage of undecided teachers was observed in item number 20 where they were asked about the relevance of generating several models in order to identify their limitations as representations of reality. The responses with the highest mean value in both countries were observed in item 19,

related to the scope of a model and whether a model can represent a specific part of a phenomenon rather than representing the entire phenomenon.

Purpose of Models: This dimension referred to teachers' knowledge of the role of models as generative tools that allow people to explain and predict phenomena. Table 40 shows the descriptive statistics for the items related to the purpose of models.

Table 40

Descriptive Statistics for QALMBT-Epistemic for Items Related to Purpose of Models

Statement		1	2	3	4	5	*M*	*SD*
4. Models can be designed with a main purpose of making testable predictions between variables.	Chile	1.9	4.6	13.2	37.6	42.7	4.1	0.9
	Canada	0	7.0	14.0	41.9	37.2	4.1	0.9
8. Models are used to help formulate ideas and theories about scientific events, phenomena or objects.	Chile	1.6	3.0	4.3	31.5	59.7	4.4	0.8
	Canada	0	0	2.3	41.9	55.8	4.5	0.5
15. Models can serve to transfer findings to another phenomenon.	Chile	2.2	2.2	15.9	40.9	39.0	4.1	0.9
	Canada	0	2.3	4.6	55.8	37.2	4.3	0.7
16. Models can be used to test assumptions about something.	Chile	1.1	4.0	8.3	44.1	42.5	4.2	0.8
	Canada	0	2.3	4.7	44.2	48.8	4.4	0.7

Note: The first row indicates the scale included in the questionnaire in which 1 = Strongly disagree; 2 = Disagree; 3 = Undecided; 4 = Agree; 5 = Strongly disagree. *M* = Mean; *SD* = Standard deviation.

The distribution of scores for this dimension concentrated between the agree and strongly agree options with a mean value of over 4 for each item. Item 8, in which ISTs were asked about the function of models to help formulate ideas or theories about something, had the most strongly agreed upon response (over 90%). Also, the vast majority of ISTs in both countries responded agree and strongly agree on item 16. In other words, for item 16, many teachers acknowledged the importance of models to test assumptions. The utility of models to transfer findings from a specific idea to another phenomenon in item 15 also revealed a high mean score, 4.27 and 4.12

for Canada and Chile, respectively. In the case of the Chilean ISTs, 15.86% of them responded on this item as undecided. Interestingly, the majority of Canadian and Chilean ISTs acknowledged the fact that models have a predictive power (item 4) with means over 4. This is an interesting result because many studies have shown that teachers rarely recognize the role of models as tools to generate information and make predictions (see, for example, Justi & Gilbert, 2002; Van Driel & Verloop, 2002).

Testing Models: This theoretical dimension referred to teachers' knowledge of the role of models to test new ideas and hypotheses. Table 41 shows the results from the QALMBT-Epistemic for the items related to testing models.

Table 41

Descriptive Statistics for QALMBT-Epistemic for Items Related to Testing Models

Statement		1	2	3	4	5	*M*	*SD*
2. Hypotheses can be tested by using a model.	Chile	2.7	5.4	13.7	41.7	36.6	4.0	1.0
	Canada	2.3	4.7	7.0	46.5	39.5	4.2	0.9
10. A model can be tested conceptually or non-experimentally.	Chile	4.8	9.7	23.1	29.0	33.3	3.8	1.2
	Canada	0	7.0	20.9	46.5	25.6	3.9	0.9
12. Models are developed to allow us to raise new questions and create new problems.	Chile	3.8	7.8	16.1	34.7	37.6	3.9	1.1
	Canada	0	11.6	9.3	39.5	39.3	4.1	1.0
17. Models need to be assessed to test their validity and fit with reality.	Chile	1.1	1.9	14.0	39.0	44.1	4.5	0.7
	Canada	0	4.7	2.3	32.6	60.5	4.5	0.8

Note: The first row indicates the scale included in the questionnaire in which 1 = Strongly disagree; 2 = Disagree; 3 = Undecided; 4 = Agree; 5 = Strongly disagree. *M* = Mean; *SD* = Standard deviation.

Overall, the items included in this dimension of testing models showed a mean of approximately 4. Nevertheless, for item 10, the mean was lower in the case of the Chilean ISTs ($M = 3.8$). That is, almost one out of four of the Chilean ISTs answered undecided on the statement that a model can be tested conceptually. A similar percentage of undecided ISTs were observed in the

Canadian sample on item 10 but the percentage of teachers who endorsed their answers as strongly disagree or disagree was just half of the Chilean sample (6.98% and 14.5%, respectively). Surprisingly, the vast majority of the teachers in both countries recognized that models need to be assessed to test their validity and fit with reality (item 17). Also, eight out of ten teachers in both countries understood that models are useful to test hypotheses (item 2).

Changing Models: This dimension related to teachers' knowledge about one of the characteristics of models in science; when the current model does not properly explain or predict a target object or phenomenon, a process of revision and modification of initial models can occur. Table 42 displays the results for the items related to changing models.

Table 42

Descriptive Statistics for QALMBT-Epistemic for Items Related to Changing Models

Statement		1	2	3	4	5	*M*	*SD*
3. Models need to be refined based on an iterative process in which empirical data compels the revision of the model.	Chile	0.8	2.2	6.5	33.1	57.5	4.4	0.8
	Canada	0	0	7.0	37.2	55.8	4.5	0.6
5. Models could be changed when deduced hypotheses do not explain the original event, phenomena or object.	Chile	2.2	6.2	12.1	32.0	47.6	4.2	1.0
	Canada	0	2.3	11.6	32.6	53.5	4.4	0.8
9. A model can be adjusted to reflect new findings.	Chile	0.8	2.7	11.0	32.0	53.5	4.3	0.8
	Canada	0	0	2.3	34.9	62.8	4.6	0.5
11. A characteristic of models is that they can be disproved when problems with its explanatory adequacy are identified by scientists.	Chile	1.3	5.7	12.1	31.2	49.7	4.2	1.0
	Canada	0	7.0	16.3	41.9	34.9	4.0	0.9

Note: The first row indicates the scale included in the questionnaire in which 1 = Strongly disagree; 2 = Disagree; 3 = Undecided; 4 = Agree; 5 = Strongly disagree. *M* = Mean; *SD* = Standard deviation.

The mean of ISTs' scores across the individual items showed that teachers had a sophisticated knowledge for the items related to changing models. For each item, the mean was over 4. For item 9, the mean value was high in both countries. For example, more than 90% of the Canadian

ISTs were agreed or strongly agreed about this statement, and, in Chile, almost 85% chose these options. This result means that many ISTs identified that a model can be refined or adjusted to reflect new findings. A large percentage of teachers also acknowledged that models can be revised through an iterative process of enrichment and revision based on the analysis of empirical data (item 3) (M_{Canada} = 4.48; M_{Chile} = 4.4); models can be changed when they do not explain the target phenomenon or hypotheses (item 5) (M_{Canada} = 4.4; M_{Chile} = 4.2); and models can be disproved when they have problems with its explanatory adequacy (item 11). (M_{Canada} = 4.0; M_{Chile} = 4.2).

In general, the main results of the QALMBT-Epistemic questionnaire revealed that Canadian and Chilean teachers had a sophisticated knowledge for the items included in the questionnaire since the majority of their means were close to 4 (agree). Chilean ISTs showed slightly lower mean values for some of the items in comparison to the Canadian ISTs (see items related to testing models). In the case of the Chilean sample, only a few items had means lower than 4. For example, item 6 ("A model differs in some degree from the reality.") showed a mean of 3.7, and more than 60% of the participants answered they agreed or strongly agreed with the statement. In this vein, it is possible that the remaining teachers do not know that models are not perfect and can be revised and modified to try to explain and represent the reality. Regarding the items related to testing models, Chilean ISTs knew less than the Canadian sample, however, the difference was not too large (see for example mean values for item 10 in Table 41). Surprisingly, a large percentage of teachers showed undecided for the questions related to the idea that models can be tested conceptually and the fact that models are developed to allow us to raise new questions and create new problems. This result might reveal that teachers do not acknowledge

the fact that models can be used as research tools which might also influence the way that ISTs use models in the classroom and how they encourage their students to reason with them.

4.4 The Relationship between IST' Knowledge of Models and Modeling and their Assessment Literacy

As mentioned previously, to address the first research question concerning whether ISTs' knowledge of models and modeling was related to ISTs' assessment literacy in MBT, I developed and administered the QALMBT questionnaire. In order to determine ISTs' assessment literacy in MBT, the QALMBT questionnaire that was administrated was comprised of three main sections; i) demographic information; ii) QALMBT-Generic and QALMBT-Modeling where the same questions were asked from a generic approach to teaching science and from a model-based teaching approach, and iii) QALMBT-Epistemic. A total of 415 ISTs from Canada (43) and Chile (372) responded to the questionnaire as shared. Table 43 shows the descriptive statistics of the outcome variable for the QALMBT,

Table 43

Descriptive Statistics of ISTs' QALMBT Score for the Canadian and Chilean Sample

Version	*N*	Mean	SD	Min.	Max.
QALMBT-Generic $_{Canada}$	43	138.1	15.8	101	169
QALMBT-Modeling $_{Canada}$	43	118.9	26.7	38	161
QALMBT-Epistemic $_{Canada}$	43	85.7	8.8	67	99
QALMBT-Generic $_{Chile}$	386	145.7	18.7	53	175
QALMBT-Modeling $_{Chile}$	386	132.3	24.2	45	175
QALMBT-Epistemic $_{Chile}$	372	83.3	10.0	27	100

Note: SD = Standard deviation.

To explore the relationship between the two scores (QALBMT-Generic and QALMBT-Modeling) and ISTs' assessment literacy, I conducted several OLS regressions with different

predictors. I used sequential regression where different models were compared in order to identify what predictors better fit the data (e.g., years of experience teaching science, knowledge of models and modeling). Below, I detail the findings from the analysis of the QALMBT by conducting regression models for each section of the questionnaire (Generic and Modeling) for each context in which the instrument was administered.

For each of the regression models, it was found that ISTs' understanding of models and modeling (QALMBT-Epistemic) was positively and significantly related to ISTs' assessment literacy. Indeed, the predictor QALMBT-Epistemic was a significant predictor of ISTs' QALMBT-Generic and QALMBT-Modeling scores in the Chilean sample of ISTs (see Tables 44 and 45). In the case of the Canadian sample, this predictor was also positively correlated, but it was not significant when the predictor topic_science_course (number of topics in science courses learned in their teacher education program) was included in the regression model. Tables 47 and 48 provide details of the regression models. First, I present the equations for each linear regression with the predictors of interest in this study, and then I present a table to compare the models with only one predictor (Model 0 in which only QALMBT-Epistemic is included) and a second model with other predictors that might better explain the data. In the case of the Chilean sample, the final and simplest assessment literacy models that best fit the data and that explain a greater percentage of the variance for the QALMBT-Generic (Equation 2) and QALMBT-Modeling (Equation 3), were found to be, respectively:

$$QALMBT\text{-}Generic_{Chile} = 136.28 + 0.26(QALMBT\text{-}Epistemic) + 2.80(assessment\_courses) + 1.60(topic\_science\_course) + \varepsilon \quad (2)$$

$$QALMBT\text{-}Modeling_{Chile} = 121.48 + 0.55(QALMBT\text{-}Epistemic) + 3.83(assessment\_courses) + 1.46(topic\_science\_course) + 0.41(year\_experience) + \varepsilon \quad (3)$$

From the equations indicated above, in the case of the QALMBT-Generic, the predictors QALMBT-Epistemic, topic_science_course, and assessment course were significant predictors of ISTs' QALMBT_Generic scores. By including these predictors, the model explained 9% of the variance in the data. Table 44 shows the different models that were compared. From the table it can be seen that Model 1 was the model that best fit the data and accounted for 8.2% of the variance. In other words, Model 1 explained 6.2% more variance in the outcome variable than Model 0. The results of this regression suggested that the QALMBT-Generic is expected to increase by 0.26 units for each unit of increase of QALMBT-Epistemic, when all else is held constant. Table 45 shows the regression model for assessment literacy in MBT or QALMBT-Modeling. From the regression model it was found that the predictors QALMBT-Epistemic, topic_science_course, assessment_courses, and year_experience_science, were significantly positively correlated to ISTs' QALMBT-Modeling scores. The predictors in the Model 1 explained 11% of the variance in the data in comparison to Model 0 with only one predictor which accounted for 5% of the variation in the outcome variable. In other words, Model 1 explained 6.1% more variance in relation to Model 0. Moreover, the results of the regression suggested that QALMBT-Modeling was expected to increase by 0.557 units for each unit of increase of QALMBT-Epistemic when all else held constant

Table 44

Models for QALMBT-Generic Score (Chile)

	Model 0			Model 1		
Predictor	Estimate	s. e.	*p*	Estimate	s. e.	*p*
Intercept β0	145.66***	0.965	.000	136.28***	2.074	.000
QALMBT-Epistemic	0.28**	0.096	.003	0.26**	0.093	.005
assessment_course				2.80**	0.991	.005
topic_science_course				1.60***	0.425	.000
Multiple R-squared	.022			.090		
Adjusted R-squared	.020			.082		

Note: s. e. = Standard error. *** denotes **α** significance level at .001, ** at .01, * at .05. The model with only one predictor (model 0) and the model that best describes the data is included in the table (model 1).

Table 45

Models for QALMBT-Modeling Score (Chile)

	Model 0			Model 1		
Predictor	Estimate	s. e.	*p*	Estimate	s. e.	*p*
Intercept β0	132.08***	1.226	.000	121.48***	2.644	.000
QALMBT-Epistemic	0.557***	0.122	.000	0.55***	0.118	.000
assessment_course				3.83**	1.216	.002
topic_science_course				1.46**	0.541	.007
year_experience				0.41*	0.167	.016
Multiple R-squared	.05			.12		
Adjusted R-squared	.05			.11		

Note: s. e. = Standard error. *** denotes **α** significance level at .001, ** at .01, * at .05. The model with only one predictor (model 0) and the model that best describes the data is included in the table (model 1).

In the case of the Canadian sample, the results from the questionnaire followed a similar trend to the Chilean sample. That is, ISTs showed higher scores for the QALMBT-Generic in comparison to the QALMBT-Modeling (see the values of β_0). In other words, ISTs' self-report about their assessment practices indicated that participants more often implemented assessment strategies when assuming a generic approach to teaching science (QLMBT-Generic) than in an approach that involved the enactment of modeling practices (QALMBT-Modeling). The process of model comparisons of the OLS regressions showed that the QALMBT-Epistemic was also positively

related to ISTs' QALMBT-Generic and QALMBT-Modeling scores, but this predictor was only significant for the QALMBT-Generic. This result might be influenced by the small sample in the case of Canadian ISTs which might have reduced statistical power. The simplest models that best fit the data and that explained more percentage of the variance for the QALMBT-Generic (Equation 4) and QALMBT-Modeling (Equation 5) were respectively:

$$QALMBT\text{-}Generic_{Canada} = 138.14 + 0.60(QALMBT\text{-}Epistemic) + \varepsilon \quad (4)$$

$$QALMBT\text{-}Modeling_{Canada} = 101.83 + 0.71(QALMBT\text{-}Epistemic) + 4.36(topic\_science\_course) + \varepsilon \quad (5)$$

The equation above revealed that in the case of the QALMBT-Generic, the model with only QALMBT-Epistemic as a predictor better fit the data and accounted for 9% of the variation in the outcome variable (see Table 46). Also, as mentioned previously, this predictor was significantly related to IST's assessment literacy in terms of their general approach to teaching science. The results of this regression suggested that the QALMBT-Generic was expected to increase by 0.6 units for each unit of increase of QALMBT-Epistemic, when all else is held constant. Table 46 shows the regression model for the QALMBT-Generic score.

Table 46

Models for QALMBT-Generic Score (Canada)

	Model 0		
Predictor	Estimate	s.e.	*p*
Intercept β0	138.14***	2.306	.000
QALMBT-Epistemic	0.60*	0.263	.028
Multiple R-squared	0.11		
Adjusted R-squared	0.09		

Note: s. e. = Standard error. *** denotes **α** significance level at .001, ** at .01, * at .05. The model with only one predictor (QALMBT-Epistemic) showed the better fit to the data.

Regarding the QALMBT-Modeling, table 47 below reveals that the predictor QALMBT-Epistemic was not significant but it was positively related to QALMBT-Modeling. It was found that for each unit of increase of the QALMBT-Epistemic, the QALMBT-Modeling was expected to increase by 0.71 units. The model that better fit the data (model 1) showed that the predictor topic_science_course was significantly related to the dependent variable and showed that for one unit of increase of this variable, the outcome variable was expected to increase by 4.358 units. This model accounted for 24% of the variation in the outcome variable in comparison to Model 0 that only accounted for 13% of the variance. In this sense, Model 1 explained 11% more of the variance in comparison to Model 0.

Table 47

Models for QALMBT-Modeling Score (Canada)

	Model 0			Model 1		
Predictor	Estimate	s. e.	*p*	Estimate	s. e.	*p*
Intercept β0	118.86***	3.81	.000	101.83***	7.428	.000
QALMBT-Epistemic	1.16*	0.44	.011	0.71	0.443	.117
topic_science_course				4.36*	1.668	.012
Multiple R-squared	.15			.27		
Adjusted R-squared	.13			.24		

Note: s. e. = Standard error. *** denotes **α** significance level at .001, ** at .01, * at .05.

To summarize, each of the above tables showed the relationship between a set of independent variables (e. g., year of teacher experience) and the dependent variable (ISTs' assessment literacy in MBT). The first research question of this study aimed to identify the variables that might predict ISTs' ALMBT, and particularly emphasized the teachers' knowledge of models and modeling as one of the main predictors. The results suggest that when ISTs know more about the nature and purpose of models, it is likely that their assessment practices related to MBT occur with more frequency in their pedagogy. Interestingly, the predictor related to knowledge of the

nature of models was also related to a more general approach of teaching (see results from the regression models for the QALMBT-Generic). In this respect, it can be claimed that if teachers know more about models, they might also know about other topics that are important in science education (e.g., the nature of science). For example, they might also have better knowledge about how to identify and assess students' knowledge and reasoning in science, and therefore, they might be able to more frequently assess students in a general approach. It is worth noting that even though the predictor related to knowledge of models was not significant in the regression model for the Canadian sample, it was still positively related to the variable. Due to the small sample size in the Canadian context, it might be possible that the statistical power was reduced (Serdar et al., 2021). Moreover, as suggested by Xu and Brown (2016), assessment literacy also is related to ISTs' identities and sociocultural contexts of teaching and learning which might explain the differences between the Canadian and Chilean sample. In the case of MBT for the Chilean sample, years of experience and the number of courses taken in assessment were predictors that were also related to ISTs' assessment literacy in MBT. In the case of years of teaching experience, this predictor is supported by Mertler's (2004) study that showed that in-service teachers score higher in the Classroom Assessment Literacy Inventory (CALI) in comparison to teachers with less experience, such as pre-service teachers. Similar results were found in Bandele and Oluwatayo's (2013) study in which the authors after administering a multiple-choice test of knowledge of assessment techniques identified that teachers with more years of teaching experience scored higher in comparison to science teachers with less years of teaching experience. In this vein, Pophan (2009) stresses that teachers need time to become assessment literate. Based on these results, I assume that teachers who have more teaching experience, will have more opportunities to practice, enrich and expand their assessment

repertoires. Another assumption is related to the number of courses taken in assessment in teaching education programs. Teachers who have taken more courses in assessment, will have more sophisticated knowledge and repertoire to assess their students. Finally, based on the results of the regression models, it can be determined that teachers with a better knowledge of the epistemology of models report they engage their students more frequently in the assessment of models. These results appear to be coherent with Justi's (2009) contention that teachers' pedagogy and their modeling practices in the science classroom are shaped by their knowledge of models and modeling. Moreover, Jones and Moreland (2005) highlight that teachers' pedagogical content knowledge influences their capacity to implement formative strategies to help students progress towards more a complex understanding of scientific ideas. In other words, ISTs hardly will assess models and modeling practices if they do not know what modeling involves in the science classroom.

To explore whether ISTs' knowledge of models and modeling and the other predictors were related to the frequency (how often) and type of assessment practices (variety of assessment practices related to each dimension included in Figure 1), the following section shows the results from the qualitative analysis related to the phase of identification and development of assessment literacy in MBT after attending an online professional development course.

4.5 MBT Assessment Practices

To further complement findings on the baseline assessment literacy of ISTs, I observed the actual classroom assessment practices of 5 Chilean ISTs to answer the Research Question 2 related to in what ways ISTs' assessment literacy about models and modeling influenced their

pedagogy. In this section, I present the assessment practices observed in ISTs' pedagogy and discuss IST's purposes when assessing students' reasoning with a model. The data sources that were analyzed qualitatively and quantitatively include: transcripts from class observations (approximately 80 minutes for each class), interviews (1 hour for each cycle), and teachers' artifacts such as administered exams. Considering that the main goal of this study is assessment, each of the data sources were triangulated in an attempt to portray how teachers included their knowledge of assessment for each of the theoretical dimensions associated with ALMBT in this study (see Fig. 1).

As mentioned previously, assessment literacy in this study assumes that ISTs possess a specific body of knowledge. This assumption is based on the integration of several theoretical dimensions related to ISTs' knowledge concerning what assessment strategies to apply, how to apply those assessment strategies, and when to implement assessment strategies in the science classroom. Note that this research also assumes that ISTs' assessment literacy might vary from a generic approach to teaching science versus an MBT approach. This distinction is important because I hypothesize that ISTs' assessment repertoire might be different when teachers assess their students after conveying content information in a generic manner compared to when they are teaching with models. In order to further explore ISTs' assessment literacy in MBT and to answer the second research question, five secondary chemistry and biology ISTs enrolled in an online professional development course (OPDC). Before and after changes to their actual classroom practices were ascertained to investigate how ISTs' assessment literacy in MBT was shaped after attending the OPDC. As it was already mentioned in the methodology chapter, this study used an explanatory sequential research design. In the phase of identification and

characterization of ISTs' assessment literacy in MBT, the qualitative data was used to assist in explaining and interpreting the findings of the phase of baseline of ISTs' assessment literacy in MBT (ALMBT) through the administration of the QALMBT. The results of the EFA in the QALMBT-Modeling revealed three factors related to ISTs' assessment literacy called i) implementation of strategies to promote the elicitation and assessment of students' models; ii) self and peer assessment of generated models, and iii) communication of assessment criteria to assess students' models. Moreover, the regression analysis revealed that ISTs' i) knowledge of the nature models and modeling; ii) the number of courses taken in assessment; iii) the knowledge of topics in science education; and iv) the years of teaching experience were predictors that were significantly and positively related to ISTs' ALMBT. In this vein, the results showed in the following section from the qualitative analysis elaborate on the results obtained from the QALMBT to refine and extend the factors and predictors identified in the phase of baseline of ALMBT. Specifically, ISTs' frequency and type of assessment practices were explored before and after attending an online professional development course (OPDC) in MBT to identify how IST's assessed students' models in the science classroom.

4.5.1 Theme 1: Implementation of Strategies to Promote the Elicitation and Assessment of Students' Models

In my conceptualization of assessment literacy, Xu and Brown (2016) was used as a framework of ALMBT (See Figure 1). In my study, the results from the exploratory factor analysis uncovered statistically three factors, that were further illustrated in the observation of ISTs' practice. Due to the fact that I used thematic analysis to analyze the qualitative data, in the phase of identification and development of ISTs' ALMBT, I refer to the uncovered factors as themes.

The first theme, related to the implementation of strategies to promote the elicitation and assessment of students' models, was comprised of six out of eight of the theoretical dimensions of ALMBT including: i) disciplinary knowledge and PCK, ii) knowledge of assessment purpose, content and methods, iii) knowledge of feedback, iv) knowledge of interpretation and communication, v) knowledge of assessment ethics, and vi) knowledge of scaffolding and learning progressions. This theme was related to the variety of assessment strategies i) to design assessment instruments, ii) to gather information through the assessment of students' models and modeling practices, and to interpret and communicate students' progress when thinking with models. Even though in some cases ISTs' assessment practices might be seen as teaching practices implemented by teachers when interacting with their students, I argue that they correspond to assessment maneuvers that teachers used to assess student achievement and reasoning with a model. Analysis and discussion of the results are provided below for each theoretical dimension.

4.5.1.1 ISTs Scarcely Use Their Disciplinary Knowledge About the Epistemology of Models and Modeling and PCK to Shape Their Instruction and Assessment Strategies in MBT

The theoretical dimension of disciplinary knowledge and PCK (as detailed in the knowledge base in Figure 1) in this study referred to ISTs' knowledge of models and pedagogical content knowledge about how to teach a science topic in particular with models. The analysis of interviews and class observations revealed that this theoretical dimension was comprised of three sub-themes. These sub-themes were not found in the statistical analysis since the Exploratory Factor Analysis (EFA) only revealed three main factors which were interpreted as main themes

in the thematic analysis. Figure 20 details the strategies related to disciplinary knowledge and PCK used by ISTs to help students elicit and assess their models. Each sub-theme is then detailed in the following paragraphs.

Figure 20

Assessment Practices for each Sub-Theme Related to Disciplinary Knowledge and PCK

Disciplinary Knowledge and PCK

a) Generating models	b) Conveying content information and curricular models	c) ISTs' epistemological knowledge of models in science
i) Asks students to create a model. ii) Asks students to analyze data to generate a model. iii) Conducts driving questions to encourage students to generate models.	i) Uses analogies to clarify ideas. ii) Provides content information about a core idea and a curricular model. iii) Contextualizes a model to motivate students. iv) Uses historical models to convey content information. v) Shows a curricular model or IST generates a model or scheme.	i) Historical models. ii) Multiplicity of models. iii) Nature and purpose of models.

The first sub-theme included a) Generating models. The three main assessment practices in this sub-theme observed among participants were i) asking students to create a model; ii) asking students to analyze data in order to generate a model; and iii) conducting driving questions to encourage students to generate models. Each of the assessment practices identified in ISTs' pedagogy is detailed below in Table 48 which summarizes the frequency and type of each assessment strategy identified by implementing the R-ASMM. Table 48 shows that three types of assessment practices were identified for the first sub-theme related to the generation of models. As it was mentioned earlier in the methodology chapter, the heat map shows the overall trend for each assessment practice ranging from 1 class to 4 classes. Nevertheless, because James,

Samantha and Eliana were observed in more than 4 classes, the average of each practice based on the number of classes observed is also reported in parenthesis.

Table 48

Frequency of Assessment Practices Observed for the Sub-Theme Related to the Generation of Models

	S		J		E		L		G	
- Asks students to create a model (e.g., explanation or conceptual model) or identify relationships between variables)		1 (1)		1 (1)					1 (1)	
- Ask students to analyze data in order to generate a model and identify, for example, trends and patterns.		5 (2.5)			1 (1)				1 (1)	
- Conducts driving questions to encourage students to generate models (e.g., explanations) based on their prior knowledge.	1 (1)	4 (1.3)		2 (1)	1 (1)				3 (1.5)	1 (1)

Note: The first column after each assessment practice corresponds to the frequency observed before the OPDC, whereas the second column refers to the frequency observed after the OPDC. The letters S, J, E, L and G refer to the first letter of each ISTs' name (Samantha, James, Eliana, Lisa and Gabriel, respectively). In parenthesis is shown the average of each assessment practice based on the number of classes that the action was counted in the R-ASMM. The colored boxes indicate an ISTs' frequency of assessment practices over the unit observed in red scale from white through dark red (blank space = not observed; pale red = observed in only one class; pink = observed in two classes; red = observed in three classes; and maroon = observed in four classes or more).

As it is indicated in Table 48, the majority of the practices included in this sub-theme were scarcely observed among the 5 ISTs when implementing the R-ASMM. These results were aligned with the findings from the QALMBT which showed that for the four items related to the theoretical dimension of disciplinary knowledge and PCK (items 1, 6, 20, 26), the 5 ISTs showed a low self-report. For example, on average for these 4 items, Eliana and Gabriel scored 2.25, whereas James and Lisa scored 2.75 in the 5-point scale (1 = never, 2 = very rarely; 3 = sometimes; 4 = Frequently; 5 = very frequently). Samantha was the only IST who showed a higher average (3.5) and I also observed she was the only one who engaged students in each practice after the OPDC. The questions related to; the development of models when conducting

investigations (item 6); the construction of models in laboratory activities (item 20); and the use of assessment to measure students' understanding of generated models (item 26) showed the lowest scores for each IST. Interestingly, the average in the scores from the sample of the Chilean ISTs who participated in the first phase of the study (3.84; 3.5; 3.37; 3.8 for the items 1, 6, 20, 26, respectively) was higher in comparison to the 5 ISTS who participated in the second phase of the study.

Regarding the first strategy investigated and shown in Table 48, asking students to create a model, the Rubric of Assessment Strategies in Models and Modeling (R-ASMM) showed that the 5 ISTs rarely engaged their students to generate their own models to identify relationships or explain a phenomenon. The R-ASMM showed that the generation of models was implemented only once among Samantha, James, and Gabriel. Samantha was the only teacher who engaged students in modeling practices that challenged students to generate and use a model to explain an idea. To do so, she challenged her students to evaluate periodic trends, for example, based on how the elements of the periodic table were arranged. The second strategy investigated included students' analysis of data to generate a model. The R-ASMM showed that Eliana and Gabriel included this strategy in only one of the classes observed before the OPDC. On the other hand, in the case of Samantha, this assessment practice was not observed before the OPDC, but I observed that she used this strategy five times in total with an average of 2.5 times after attending the OPDC (specifically in the first and second class after the OPDC). For example, in the first class after the OPDC, Samantha asked her students to analyze the electron configuration of the elements and based on that, students were asked to create a model of the periodic table and suggest criteria to organize the elements. This example is analyzed in more detail later when I

present the theme related to the engagement of students in peer and self-assessment. A third strategy was related to formulating driving questions to help *students generate a model*. The extant literature on MBT have consistently shown that formulating a driving question is a common practice that ISTs implement to judge students reasoning (see for example, Campbell et al., 2012; Ke & Schwarz, 2019; Lamar et al., 2018; Louca & Zacharias. 2015; Ryu et al., 2015). This strategy was observed in both before and after the OPDC in the case of Samantha (1 time before the OPDC and 4 times with an average per class of 1.3 times after the OPDC) and Gabriel (3 times before the OPDC with an average per class of 1.5 and 1 time after the OPDC). Samantha and Gabriel formulated questions to identify students' initial ideas about a model. Table 49 shows recorded examples of driving questions formulated by ISTs to help students generate their models.

Table 49

Examples of Driving Questions by ISTs to Help Students Elicit their Initial Ideas

IST	Lesson	Example
S	5 (Pre)	T: How do you imagine a sheet of metal, for example, a sheet of copper? How do you imagine it at the atomic level? S2: With many atoms attached.
G	1 (Pre)	T: Galapagos is in Ecuador. It is an archipelago of islands. They are different types of finches. What differences do you see in the finches? (The teacher shows an image with different species of finches) S3: they have evolved. T: No, I'm saying it, looking at it. S4: The beak shape. T:. (…) What is the shape of the beak?

Note: The second column indicates the number of the lesson. Pre= before the OPDC; Post= After.

The Table 49 above shows examples of driving questions that teachers were observed using as actions to assess students' initial ideas about a model. The analysis of the interview before the OPDC also revealed that ISTs rarely asked their students to generate models. For example, in

Gabriel's interview before the OPDC, when he was asked about engaging students in the generation of models, he stated, "No, I think I do it rarely (…) because of the time, not enough time, and also because I think that I have not understood it so much like that.". In the contrasting case of James, he indicated in the first interview that he used the generation of models, on the other hand, as a common practice when assessing students through exams. He appeared to ask students to generate a model that he had already taught in class rather than asking them to generate a new model based on the students' own initial ideas. He stated, in the case of written assessment, "Students have to make some representation of some phenomenon, process or structure that we have seen in class." In another example of a summative assessment, James similarly stated "In the last test, for example, we asked students to represent a model of the peripheral nervous system. It was a model that we had seen in class." From this excerpt, it can be noted from the underlined text that James assessed the generation of a model based on how students remembered a curricular model studied in class instead of asking students to generate a model by themselves in order to explain a phenomenon. In terms of the OPDC, the analysis of the interviews showed that several ISTs enriched their ideas about how to challenge students to generate a model after the OPDC. For example, in the second interview, James acknowledged an important aspect of the generation of models is to help him identify how students enrich their ideas about a model. He stated, "[T]he fact that they can develop versions of their own model gives me a lot of information regarding the level of knowledge they could have developed and the depth they have. I have a lot more information to work with." Similarly, Samantha said in the interview after the OPDC:

> Yes, it (pedagogy) has changed (…). I have been more explicit with the students about the fact they have to build models. I have asked them to manipulate models in such a way that they can do targeted actions and better understand what they are doing.

The interview showed that Samantha's ideas about the role of models and modeling and how to engage students in using models changed after attending the OPDC. Specifically, she acknowledged that models can be manipulated to allow us to explain target phenomena, which had not been acknowledged before in a similar question in the interview before the OPDC. In conclusion, the results from the R-ASMM, interviews and QALMBT, showed that ISTs did not often implement strategies to assess the generation of models (on average once per class). Moreover, the type of strategies used in the classroom to promote the elicitation of models occurred mostly through driving questions which might show the limited repertoire that teachers have in terms of PCK when teaching with models. Nevertheless, it seems that after the OPDC ISTs started to understand the importance of explicitly encouraging students to generate models in the science classroom, as it was identified in Samantha's example during the interview.

The second sub-theme that was observed among ISTs was related to b) Conveying content information and curricular models. Four main strategies were identified in ISTs' pedagogy based on an analysis of class observations. These strategies included: i) provide content information about a core idea or ii) a curricular model, iii) show a curricular model, and iv) ISTs generating a model or scheme. Table 50 shows the assessment practices observed in ISTs' pedagogy.

Table 50

Frequency of Teacher Assessment Practices Observed for the Sub-Theme Related to Convey Content Information and Curricular Models

	S		J		E		L		G	
- Provides field-specific bodies of knowledge from a more generic perspective (e.g., definitions, basic rules/calculation)	5 (2.5)	7 (1.75)	16 (4)		4 (4)	1 (1)		4 (2)	4 (4)	
- Provides content information in any form about a "curricular model". Explains a mechanism.	2 (2)	18 (4.5)	29 (5.8)*	15 (3.8)	17 (5.7)	3 (3)	15 (3.8)	3 (1.5)	4 (4)	5 (2.5)
- Provides a representation (e.g., image/video) of the expected curricular model without encouraging the construction of the model.		2 (1)	8 (1.6)*	8 (2.7)	1 (1)		10 (2.5)	6 (3)	1 (1)	
- Generates a scheme/drawing to represent a target (phenomenon/object/mechanism).	1 (1)	2 (1)	7 (2.3)	3 (1.5)	1 (1)	1 (1)	3 (3)			

Note: The letters S, J, E, L and G refers to the first letter of each ISTs' name (Samantha, James, Eliana, Lisa and Gabriel, respectively). Colors: blank space = not observed; pale red = observed in only one class; pink = observed in two classes; red = observed in three classes; and maroon = observed in four classes or more. *: indicates that the action was observed in 5 classes.

From Table 50 it can be observed that providing content about a curricular model by lecturing students about content (e. g., defining the process of exocytosis) was the most common strategy used by teachers to teach the models that they would assess in class. This strategy was observed across the five ISTs before and after the OPDC. The R-ASMM showed that James was the teacher who most often used this strategy with an average of 5.8 times per class before the OPDC and 3.75 times per class after the OPDC. ISTs typically introduced a new idea or content by providing definitions or by lecturing about disciplinary core ideas related to describe a model, system, mechanism, or a phenomenon. ISTs' often provided field-specific bodies of knowledge using a lectured-based perspective. Samantha used this strategy more frequently after the OPDC (2 times before the OPDC versus 18 times after the OPDC with an average per class of 4.5 times). This difference in Samantha's instruction compared to the observations before the OPDC

might be explained by the following reasons. Firstly, the observed classes that Samantha taught before the OPDC were mostly oriented towards algorithmic-problem solving and teaching students how to determine the concentration of an analyte sample. Therefore, models were not explicitly taught to students even though a chemical equation can be treated like a model. Nevertheless, based on what I observed on Samantha's pedagogy before the OPDC, she focused on the resolution of mathematical problems instead of emphasizing the visualization of chemical equations using, for example, particulate models. Secondly, in the class observations after the OPDC, Samantha introduced the topic of periodic trends based on students' prior ideas of the atomic model. In this sense, Samantha appeared to more explicitly link the use of models with the unit she was currently teaching after the OPDC. Eliana and Lisa also used this strategy (provide content information about a curricular model) before and after the OPDC, but this strategy occurred four times more often before the OPDC (more than 15 instances before the OPDC and 3 times after the OPDC). The difference in the frequency might be because a higher number of classes were observed before the OPDC which can be identified when the same practice is analyzed in terms of average. For example, Eliana provided content information, on average, 5.6 times per class before the OPDC (3 classes in total), whereas after the OPDC occurred three times in one class. Similarly, in Gabriel's case, there was not a significant change in the use of this strategy that was observed four times before the OPDC in one of his classes (average 4 per class) and 5 times across two classes after the OPDC (average 2.5 per class). Table 51 shows some illustrative examples of how teachers conveyed content information while teaching models. The examples corresponded to class observations that were videorecorded and then transcribed.

Table 51

Examples of Practices Identified in ISTs' Pedagogy Related to Content Information Model

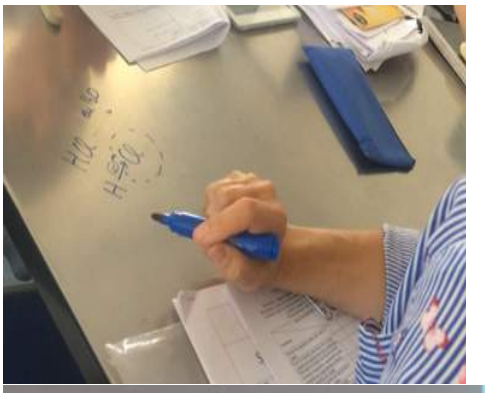
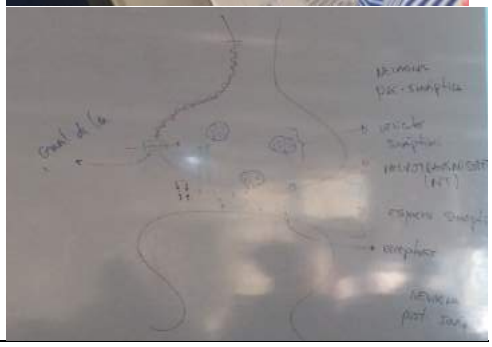
IST	Lesson	Example
J	1 (Before OPDC)	T: The exocytosis process occurs when this vesicle approaches the plasma membrane, and as it appears here, it fuses with it, releasing what it has inside.
L	1 (Before OPDC)	This enzyme called rubisco or RuBP, is the one that starts the cycle, captures carbon and goes through various reactions until glucose is synthesized and again, releases this glucose, and again reaches the beginning where it captures more oxygen again.

In these examples from Table 51, we see ISTs describing specific content information when teaching with models. In the case of James (see J above in Table 51), in the first class before the OPDC, he defined exocytosis after asking students to define the concept based on their prior knowledge. I observed that he assessed students' prior knowledge with an initial question; however, he provided the definition after not receiving an answer from students. Similarly, Lisa (see L in Table 51) also provided content information with the purpose of teaching the target model that students were expected to learn in class. Another strategy that was observed in each IST's class was related to the representation of the expected/target curricular model without encouraging the construction of a model by the student. The R-ASMM detected that ISTs often used diagrams or schemes, such as conceptual maps, to introduce the expected curricular model that students must learn. Visual representations were a common resource that teachers used before and after the OPDC to teach students the expected curricular model. This practice was commonly used by James (8 times before and 8 times after the OPDC) and Lisa (10 times before the OPDC and 6 times after the OPDC). To a lesser extent, Samantha, Eliana and Gabriel were also observed using this strategy. The last strategy that teachers used to convey content information and curricular models included the generation of a scheme or diagram by the teacher

to represent a target model. Table 52 provides data on diagrams generated by the teachers and not the students.

Table 52

Examples of Drawings Generated by the Teacher to Provide Content Information About a Model

IST	Lesson	Example	
S	5 (After OPDC)	S2: Teacher what is a strong acid? T: When you have an acid, for example, hydrochloric acid, and you put it in water, inside here you have a covalent bond but it is polar (….) So, we said by electronegativity, it will pull electrons (Cl) that are in a covalent bond. Therefore, the electrons are going to be around the chlorine atom longer (teacher draws the dissociation of HCl in aqueous solution).	
J	1 (Before OPDC)	T: Pre-synaptic neuron, post-synaptic neuron, we're looking at the detail of communication at the end of one neuron and the end of another (…) I have the nerve impulse that is moving, I have all the ion exchange through the plasma membrane.	

I observed that ISTs before and after the OPDC created on a few occasions the drawings instead of asking students to generate their models by drawing, for example, a specific phenomenon. Only James was the teacher who more often generated schemes with a frequency of 7 times in three classes and an average of 2.3 times per class before the OPDC, and 3 times after the OPDC in two classes with an average of 1.5. In the examples included in the table above, James (see J in Table 52) drew a model of a presynaptic neuron and postsynaptic neuron to explain how the chemical synapse occurs. After generating the model himself, he asked his student to copy the

drawing into their notebooks with the purpose of helping students memorize the mechanism underlying the neural communication. Unlike the other ISTs, Samantha drew a model to clarify students' conceptual doubts. She asked driving questions ("[W]hich one is more electronegative in the periodic table? Chlorine or hydrogen?") to assess students' reasoning with the model generated by the teacher. It is worth mentioning that Gabriel did not use this strategy to convey content information about a model.

Based on the four strategies identified in the R-ASMM and detailed above, the results of the interviews revealed that teachers often used models to convey content information. Similar results have also been found in studies conducted by Guy-Gaytán et al. (2019), Kawasaki and Sandoval (2020), and Wilkerson et al. (2018). A common practice that was identified in the interviews before the OPDC was related to the provision of a representation of the expected curricular model, for example, through images or videos. In this example of the strategy, teachers did not ask students to generate a model; rather it was facilitated by the teacher to help students understand a concept or an idea. For example, in the interview before the OPDC, Eliana said "Chemistry is something super difficult to see because it is something super abstract. So, I always try to either draw on the whiteboard or show videos, or show animations, so that they get an idea how things work through a model or a representation." Similarly, James reported in the first interview before the OPDC:

> I review the contents with them the concepts and I show them some models that summarize the contents. I start reviewing the information with them quickly. I go step by step teaching each one of them from the parts of the unit referring to the models that

we have been studying. Therefore, they already have a mental image of the content that we will study later.

This example suggests that James often introduced models to teach students the target model. As it is exemplified above in the underlined portions in Eliana's and James' answers, it was identified that the ISTs used the strategies related to provide content information and generate schemes not only to teach a specific model but also to guide students in the study of specific models or features of a model that ISTs were interested in assessing in the science classroom. This practice related to guiding students in the elements and features to be included in a model has also been identified in Kawasaki and Sandoval's (2020) study with secondary teachers after participating in a professional development program in the United States.

The third sub-theme for the dimension of disciplinary knowledge in PCK included: c) ISTs' epistemological knowledge of models in science. This sub-theme mainly emerged from the analysis of the interviews and included four main codes related to it: i) historical models (suggests the use of models to help students understand the changing nature of scientific models), ii) multiplicity of models (acknowledges the fact that scientists use/develop multiple models), and iii) nature and purpose of models (acknowledges the relevance of the purpose and nature of models as generative tools to explain and predict phenomena). It is worth mentioning that historical models were observed in Samantha's, Lisa's and Gabriel's pedagogy when teaching with models; however, I coded them as "Historical_model_content_information" (included in the second sub-theme) since it was a practice that was observed when teachers taught content related to historical models. I made this distinction because this third sub-theme

was more related to teachers' knowledge of the epistemology of models in science rather than how they put it into practice when providing information to help students understand the context in which the historical model was developed. Samantha and Gabriel were the only teachers who emphasized the relevance of including historical models in their classes. In the interview before the OPDC, when asked about how she used assessment to help students enrich their inquiry skills when thinking with models, Samantha recognized the role of historical models to help students identify how models evolve and understand their utility and limitations. She said in the first interview:

> The best example, that I have achieved is with the Rutherford gold foil model (…) first, there is a historical transfer to recognize that at that time little technology existed. Second, to realize the knowledge that existed of the atomic structure at that historical moment, so that they take various elements very different from each other, consider the technology that existed at that time, which allowed Rutherford to design an experiment that with its results they renewed the previous model and could with these results explain more phenomena as a consequence of this atomic structure.

In the underlined portion above, Samantha appears to recognize that scientific knowledge is tentative and empirically based. Moreover, she emphasizes that this knowledge is socially and culturally embedded since scientists often use and revise prior models to identify their limitations and scope. In the case of Gabriel, his ideas about how to use historical models in the classroom appeared enriched after the OPDC since in the interview before the OPDC he did not emphasize the relevance of teaching scientific models. In the interview after the OPDC he stated:

> Historical model, which for me was not necessarily working with a model, but for me it was rather a review through the historical context of the ideas that some scientists propose (…) I try to do it in class, but I didn't necessarily see it as working with models.

In the above example, Gabriel acknowledged that historical models can be studied in class not only to teach the context in which a model was created but also to help students work with a model in order to understand its utility. For example, regarding the multiplicity of models, the analysis of the QALMBT-Epistemic showed that the 5 ISTs had a proper knowledge for the five items included in this dimension before the OPDC. On average, the 5 ISTs and the Chilean sample scored 4 points or higher for this dimension ($M_{Samantha} = 4$; $M_{James} = 4.3$; $M_{Eliana} = 4$; $M_{Lisa} = 4.3$; $M_{Gabriel} = 4.8$; $M_{Chilean\_Sample} = 4.3$). During class observations, I only observed that Samantha used this knowledge in class which might reflect that she became more assessment literate after the OPDC since she was able to use her disciplinary knowledge (knowledge of models) and PCK (how to teach models and modeling) to help students reasoning with models. For example, in the third class after the OPDC she informed the learning goals at the beginning of the class related to "Analyze the metallic character of the elements in the periodic table; Analyze nomenclature of oxides when reacting with water; Evaluate chemical reactions of elements of groups IA and VI.", she answered a student's question as follows:

> S1: Do acids are acidic [*sic*] because hydrogen dissociates in water?
>
> T: You said it very well. The definition of acids is based on three different theories, but all three look at acid in different and complementary ways. One way is that this substance, if you put it in water, will ionize and then dissociate (Teacher writes $HNO_{3(Aq)} \Leftrightarrow H_{(ac)} + NO_3^-$).

The excerpt above shows that Samantha used her knowledge about the multiplicity of models in order to answer the student' doubt. She used students' answers (Arrhenius model of acid in which acids release protons on dissociation in water) to briefly introduce the Bronsted-Lowry model in which acids are substances that can donate a proton. Similarly, Eliana, who was also a chemistry teacher, mentioned the same example which revealed her knowledge about the multiplicity of models. Eliana in the interview before the OPDC stated that she included multiple models in her pedagogy when teaching disciplinary core ideas. She emphasized the fact that in science different models can be developed to explain the same phenomenon by integrating different theories that can coexist. For example, Eliana said:

> When I teach acid base, we have different theories, right? And we make a model based on the three theories. Normally, the one that is harder for the students, the most, is Lewis, because it has valence electrons, because they don't know Lewis electrons, which is another model.

From the above examples, it can be seen that Eliana pointed out the role of models as representations that can be used to explain different theories. In the case of James, he suggested the use of multiple models to clarify a curricular model by using different representations. James, in the interview before the OPDC, said:

> Generally, when I think of some process or some phenomenon or structure that is very complex to understand immediately, I usually use more than one model. I move to another similar model, of similar complexity, but with different characteristics, or I use a simpler model.

In the underlined section above James referred to the multiplicity of models from a pedagogical approach and to their role in helping students understand an idea or concept from different perspectives.

The nature of models was another aspect that was identified among the responses to interview questions. ISTs barely mentioned the nature of models in the interview before the OPDC. Nevertheless, Eliana acknowledged in the interview before the OPDC (see underlined section below) the nature of the atomic models as representations that have evolved and that are useful to explain molecular properties. She said:

> For example, in the case of the atom, regarding the classic model that comes to mind is Rutherford's, they should not stay with that, that is reality. Because in reality, they haven't even seen an atom, because deep down that is a representation. If they keep thinking about Bohr's or Rutherford's model, then it's difficult for them to understand polarity. Because they will review concepts such as electronic clouds, hence, the electronic cloud before the quantum mechanical model did not exist.

In the interview after the OPDC, ISTs enriched their ideas about the nature of models. For example, in the interview after the OPDC, Eliana suggested that before asking students to create a model, she must help students conceptualize what models are with a black box (see underlined section below). She said:

> In the beginning, I would try to explain to them what a model is, for example, in eighth grade, they will do the black box experiment, eh ..., and there I am going to explain

> what a model is, a hypothesis, a theory. Because in that case they have to formulate a hypothesis, they have to make a representation of what they believe about what is in the box. Then, they have to build a model of what is in the box and finally see what is in the box.

In the case of James, he pointed out that one of the most important aspects learned from the OPDC was the fact that models are tentative. Interestingly, James did not explicitly acknowledge this element of the nature of models in the interview before the OPDC. In the interview after the OPDC; however, he pointed out, "The idea of evaluation (of a model) struck me the most in terms of always making a critical analysis of the model and understanding it as an imperfect version of reality." Samantha also emphasized the utility of models as constructions that we use to understand a specific component of reality. In the interview after the OPDC, she said that the OPDC made her aware of being more explicit when referring to models in her class, "I have been explicit in explaining the advantage of a model, how to build and modify it, if it has limits or not, put margins on the model, and perhaps explain some part of what they are trying to understand and not entirely." She also pointed out that her ideas about what a model is changed after the OPDC. She stated:

> Yes, it (OPDC) helped me because I was not clear on the conceptualization of a model. Many times, I thought it was like something much more elaborated, rigid, almost like a 3-D scale representation, more concrete than even a virtual 3-D model. Concrete in the hands. Therefore, if I did not have conceptualized the concept of a model and how to work with it and how to develop it, my students did not have this idea either."

The above examples from the interviews show that before the OPDC, science teachers understood models mostly as resources to communicate a pre-defined curricular model; however, after the OPDC, in the interviews, teachers acknowledged the role of models as tools that can be manipulated to explain a specific target. In this sense, these results mostly based from the interviews revealed that ISTs became more assessment literate for the dimension of PCK. Another aspect of the nature of models that was also identified in the interviews was related to ISTs' ideas about the purpose of models. Before the OPDC, ISTs indicated the role of models to explain or visualize phenomena. For example, Gabriel stated the role of simple but accurate models was to explain phenomena before the OPDC. He stated:

> A model would be something like that allows to explain the phenomenon, the model has to be understandable. A too complex model suddenly distracts you from what you are looking for the explanation of the phenomenon. In other words, if it is too limited, the model does not work.

In the case of James, before the OPDC he indicated that he used videos in his class with the purpose of describing something and helping students to "visualize the functioning of the endocrine system, in a general way". Similarly, before the OPDC, in the interview Lisa explained she used models in her pedagogy to help students understand mechanisms and identify key steps. Lisa said:

> If we are talking about mitosis, they need to understand that there is an arrangement at the equator of the cell, that they [chromosomes] separate, then travel to the poles. They need to understand these key points that occur in each step.

After the OPDC, Samantha and Lisa showed a more sophisticated understanding of the purpose of models. For example, Lisa, in the underlined portion below, acknowledged the fact that students could create models to make predictions and understand the principles of inheritance. She stated:

> I would ask them first what they think it could be obtained, on what they believe and that they see that there are genes that are dominant and recessive. And then, ask them to make the Punnet square diagram and make a comparison…Because in the end if as you do several you realize that you can predict by Mendel's laws. One can predict which are the "F" [The progeny resulting from the crossing or filial] that one will have.

After the OPDC, Samantha (see underlined portion below) also emphasized the role of models as tools to generate and communicate information and make predictions. She stated:

> because making students aware of manipulating (a model), that they can manipulate the information, that they can make predictions with a model, (…) that their models can become obsolete when faced with certain questions that one asks. They found out that they could communicate with their model.

In summary, evidence suggests that the ISTs enriched their disciplinary knowledge (e.g., nature of models, multiplicity of models, and purpose of models) and PCK (how to engage students in the generation and revision of tentative models) in MBT after the OPDC. For example, ISTs started to understand that models are not merely a representation of a target phenomenon, instead, they are a theoretical reconstruction of the original (Krell & Krüger, 2015), as suggested by the analysis of the interviews after the OPDC. Also, Samantha and Lisa, in the interviews

after the OPDC, acknowledged the fact that models are tools that are built to help us make predictions about a phenomenon of interest. This change in ISTs' knowledge of the purpose of models as research tools that help students to make predictions is in line with a body of research that has suggested that ISTs' epistemological knowledge of models and modeling helps them to design and implement strategies to identify and assess students' generated models (Tay & Yeo, 2018; Werner et al., 2019; Williams & Clement, 2019; Windschitl et al., 2008). By acknowledging the purpose of models, it is possible that ISTs might use their epistemological knowledge of models in their instruction when designing and implementing formative assessment strategies to measure students' reasoning with models.

Taken together, Table (53) summarizes the results from the Rubric of Levels of Proficiency in Assessment Literacy in MBT (R-LPAL), which was used to characterize ISTs' ALMBT and compare their pedagogy. The R-LPAL was used to represent the overall analysis of the data sources (R-ASMM, interviews, and ISTs' artifacts) for the total number of class observations. In other words, the R-LPAL compared two data points, before and after the OPDC. As shared in Chapter 3, four levels of proficiency were considered in this rubric where i) novice assessment literate teachers corresponded to those ISTs who were just getting started to familiarize themselves with how to engage and assess students in MBT for a specific indicator related to each dimension included in Figure 1; ii) advanced beginner referred to teachers who partially developed or mentioned assessment strategies in his/her pedagogy but lacked important elements related to each theoretical dimension included in Figure 1; iii) competent referred to those teachers who implemented or mentioned assessment strategies but some elements related to MBT were not included in their pedagogy (e.g. The model is generated by students and it is used

by them to explain a phenomenon but not to make a prediction. If the role of predictions is mentioned but not observed in their pedagogy, teachers are labeled as competent.), and iv) advanced assessor which corresponded to those science teachers who showed in their pedagogy (practice/enactment and declarative levels) that they were able to achieve the full description of the indicator. The following table shows the results from the R-LPAL for the theoretical dimension of disciplinary knowledge and PCK. Two major indicators were used in this dimension to summarize how teachers' disciplinary knowledge and PCK influenced their pedagogy. In the case of the R-LPAL, the heat map does not refer to the frequency in which the indicator was observed, instead it refers to ISTs' level of achievement of the indicator (aqua blue = novice; cyan blue = advanced beginner; blue = competent; dark blue = advanced assessor).

Table 53

R-LPAL for the Theoretical Dimension of Disciplinary Knowledge and PCK

	S		J		E		L		G	
- Engages students in the generation of a model which is used as a research tool to generate information and understand a mechanism or phenomenon.										
- Both explanatory and predictive power are tested by students after constructing or using a model.										

Note: The first column after each assessment practice corresponds to the frequency observed before the OPDC, whereas the second column refers to the frequency observed after the OPDC. The letters S, J, E, L and G refers to the first letter of each ISTs' name (Samantha, James, Eliana, Lisa and Gabriel, respectively). Colors: aqua blue = novice; cyan blue = advanced beginner; blue = competent; dark blue = advanced assessor.

The R-LPAL detailed above was based on the overall analysis of the data sources from interviews, class observations and IST artifacts (if applicable, for example, lesson plans, exams). Samantha was the only ISTs whose pedagogical enactment reached more complex levels of proficiency for the disciplinary knowledge theoretical dimension, as suggested by a heat map.

For example, in the first class after the OPDC she asked students to generate a model of the periodic table and manipulate it to generate information related to the periodic trends which was observed by implementing the R-ASMM and analyzing the transcription of classroom observation. In the case of James (practice/enactment level from the R-ASMM), he asked students to generate explanations in class, but he did not challenge students to use their models and make predictions, whereas, in the case of Lisa (declarative level from the interview), the R-LPAL showed that she acknowledged the explanatory and predictive power of models in the interview after the OPDC; however, she did not engage students to generate models. In both cases, James and Lisa showed a competent level since they enriched their disciplinary knowledge and PCK after the OPDC but did not achieve the advanced assessor level as it was identified in Samantha's case. Even though Eliana and Gabriel enriched their knowledge about MBT after the OPDC, they still used and acknowledged the role of models mostly as tools to provide content information. The evidence suggests that Samantha, in comparison to the remaining ISTs, had a more sophisticated knowledge of PCK which was enriched after the OPDC. As it was suggested by the linear regression model from the analysis of the questionnaire QALMBT, it is likely that years of teaching experience influenced how Samantha enacted MBT. In this sense, Samantha might have owned a repertoire of varied strategies or activities that she was able to easily access and adapt to an MBT approach. Once Samantha became more literate about this dimension of models and modeling, it is likely that she was able to modify her pedagogy by making instructional decisions that allowed her to facilitate the elicitation and assessment of models. Most previous research has indicated that teachers' teaching experience is an important factor that appears to lead to the integration among disciplinary knowledge and pedagogical knowledge (Chan & Yung, 2018; Friedrichsen et al., 2008; Grossman, 1990; Kind, 2009; Schneider &

Plasman, 2011; van Driel, 1998) which might explain why the remaining ISTs rarely promoted the elicitation of models.

4.5.1.2 Science Teachers' Purposes of Assessment Shapes how they Engage Students in Modeling

This theoretical dimension refers to the purpose of formative and summative assessment in the class to assess students' performance when working with models. The analysis of the class observations, R-ASMM and interviews reported that teachers used assessment mostly to assess rote learning and students' algorithmic problem-solving skills without challenging them to reason with a model. In line with these results, the analysis of the QALMBT-Modeling for the theoretical dimension of the purpose of assessment (items 2, 10, 15, 27, and 31) also showed that Chilean ISTs reported that they did not often implement assessment strategies for this dimension. As it was reported in the statistical results, the mean of four of the items for the total sample of ISTs that answered the QALMBT-modeling (31= 3.5; 10 = 3.8; 27= 3.6; 15= 3.7) was close to 3.5. For example, for item 31 related to the use of assessment to evaluate the internal consistency or coherence of various models constructed by a student, ISTs reported an average of 3.52 being 3 "sometimes"! and 4 "frequently". Regarding these items, the 5 ITSs also showed a low self-report. For instance, items 31 (assess the internal consistency of various models) and 27 (I assess how students make judgments in science based on reasoning with a model) showed an average of 2.8. Eliana reported an average of 1.6 for the 5 items which revealed that she used assessment from a traditional approach to teaching science since she rarely engages students in practices related to the theoretical dimension of assessment purpose, content, and methods. In other words, she uses assessment to assess the acquisition of knowledge, which was also observed in the

implementation of the R-ASMM (see Table 54). Similarly, James reported an average of 2.8 for the five items, which reinforced the results from the R-ASMM that revealed he mostly assessed students' subject matter knowledge related to a model. In the case of Gabriel and Samantha, they reported an average of 3.4 for the items. These results suggest evidence that in the case of Samantha, even though she did not engage students very often in reasoning with models before the OPDC (based on the results from the QAMLBT-Modeling and the class observation), after the OPDC, the analysis of interviews, which are detailed later, showed that she was able to challenge her students to use and think with a model. Interestingly, Lisa reported for the 5 items an average of 4.2. Nevertheless, the analysis of the R-ASMM and interviews showed that she followed a more traditional approach to use and assess models in her pedagogy which focused on assessing the knowledge of features or mechanisms related to a model. Table 54, showed below, summarizes the results from the R-ASMM based on the class observation.

Table 54

Frequency of ISTs' Strategies Observed for the Theoretical Dimension of Assessment Purpose, Content and Methods

	S		J		E		L		G	
- Emphasizes the use of algorithmic problems to help students calculate an answer or a solution to a problem.	19 (4.8)	3 (1)			11 (2.2)*	7 (3.5)				
- Formulates a question that focuses on rules or heuristic techniques which are field-specific for each discipline	3 (1.5)	4 (1)				1 (1)				

Note: The letters S, J, E, L and G refers to the first letter of each ISTs' name (Samantha, James, Eliana, Lisa and Gabriel, respectively). Colors: blank space = not observed; pale red = observed in only one class; pink = observed in two classes; red = observed in three classes; and maroon = observed in four classes or more.

From the table, it can be observed that both type of practices, judging algorithmic problem solving and assessing students' use of rules or heuristics were only observed among chemistry

teachers (Samantha and Eliana). The frequency with which Samantha emphasized the use of algorithmic- problem solving decreased after the OPDC. I observed that after the OPDC, instead of just asking students to do calculations, she asked her students to first interpret information, for example, from a table with the melting points of different elements, and then formulate an explanation about trends. This strategy stands in contrast to algorithmic problem solving in the observations before the OPDC in which she only focused on reinforcing how to do calculations. Moreover, the analysis of the interview data revealed that teachers were more specific about including models in their explanations, for example, by recognizing the steps of the GEM cycle and modeling practices after the OPDC. The first strategy included in Table 54 related to judging algorithmic problem solving to assess students reasoning was coded as the set of step-by-step procedures when solving exercises and calculating a specific solution. This strategy often occurred after the science teacher presented a new topic and explained how to solve a similar problem. Samantha (19 times before the OPDC in 4 classes with an average of 4.75 times per class, and 3 times after the OPDC in 3 classes with an average of 1 time per class) and Eliana (11 times before the OPDC in 5 classes with an average of 2.2 times, and 7 times in 2 classes after the OPDC with an average of 3.5 times) used algorithmic problem-solving in chemistry as a common practice before and after the OPDC. ISTs provided exercises (work handout) with all the required data and asked students to solve problems to assess their ability to apply their acquired knowledge. The number of instances that algorithmic problem-solving was observed after ISTs attended the OPDC dropped considerably in the case of Samantha (see second column, after the OPDC, in Table 54, “S”). Table 55 shows some examples of classroom episodes that included algorithmic problem-solving. For example, Samantha and Eliana

emphasized using algorithmic-problem solving to assess students' reasoning about the steps required to calculate a specific answer which is showed in the underlined portions.

Table 55

Examples of ISTs' Activities that Included Algorithmic-Problem Solving

IST	Lesson	Example
S	3 (Before OPDC)	T: <u>They are asking you, what is the percentage of iron (III) or (IV) right here, not that it is iron alone, iron (III) or (IV), and now we have to do the stoichiometry. For one iron (IV), how many irons are there in a molecule? How many iron atoms do you have?</u>
E	1 (After OPDC)	S2: (The student asks the teacher how to calculate the equilibrium constant for the reaction (0.21 moles) $PCl_{5\,(g)} \rightleftharpoons$ (0.32 moles) $PCl_{3(g)}$ + (0.32 moles) $Cl_{2(g)}$) T: <u>Yes, 0.21 moles of PCl_5, the exercise is giving you the moles, you have to divide it into liters and then get Kc (equilibrium constant).</u>

From the underlined examples above, it can be seen that both teachers told students how to do the calculations without asking them to suggest a solution or developing their own understanding of the exercise. Based on teacher and student interaction, I inferred that the purpose of the strategy implemented by Eliana and Samantha was to assess and reinforce algorithmic problem-solving skills studied in the class. A second assessment strategy used by the ISTs included the formulation of questions that specifically focused on field-specific rules with the purpose of teaching disciplinary core ideas but Samantha and Eliana missed opportunities to assess students' reasoning with a model, for example, when teaching stoichiometry. This action was mostly observed in Samantha before and after attending the OPDC (3 times before the OPDC in 2 classes, with an average of 1.5 times, and 4 times after the OPDC in 4 classes with an average of 1 time per class). For example, in lesson 4 before the OPDC, Samantha explained to students how to do the calculations to determine the percent yield and purity in a reaction between sodium hydroxide (NaOH) and an acid $(COOH)_2$. (e. g. "<u>What do you need to calculate the percentage of</u>

purity?" and "You have to convert them to grams, what calculation are you going to apply?") Samantha asked the student the steps needed to determine the amount of the compound. I observed she used this strategy to assess if the student understood the mathematical relationship between moles and grams, the concept of purity, and the substances involved in the reaction without using or emphasizing space-filling molecular models to illustrate the relationships.

The analysis of the interviews revealed the purpose of assessment in terms of assessment when teaching with models of the 5 ISTs. In this sub-theme related to assessing students' subject matter knowledge of a model, ISTs reinforced disciplinary core ideas studied in class before assessing them. For example, Lisa used assessments to measure students' acquisition of key ideas studied in the class. On several occasions, she used short summative exams at the end of the class. On these occasions, the assessments appeared to be given to check if students were paying attention to her class and help them identify the main ideas studied in the class that might be assessed in a summative exam at the culmination of a unit. Lisa remarked in the interview before the OPDC "I use short quizzes at the end of the class that are oriented to cover something very specific that I said. The same answer I said it before and to see if they are also attentive." In the underlined text, Lisa stated she assessed students' subject matter knowledge through short quizzes to align students' understanding of a model in her class. After the OPDC, Lisa and Eliana also emphasized they used models to measure students' acquisition of key ideas or elements taught about a model. For example, when teaching photosynthesis, Lisa stated in the interview after the OPDC:

> They should show the correct structures, that there is a correct relationship between structure and chemical reaction. That there is a direct relationship with the type of cycle

> they are seeing. With what phase, if it is light-dependent phase. In what moments it happens.

Similarly, Eliana pointed out in the interview after the OPDC that "The main objective is to extract information from what they were able to learn and what they were missing regarding the content we were studying." This excerpt above suggests that Eliana used assessment to gather information about students' understanding specifically to identify what elements of the curricular model students lacked. Interestingly, after the OPDC, James mentioned he struggled to design summative instruments to assess the acquisition of a curricular model. In the second interview, he stated:

> When I was thinking about GEM cycle, I had always thought about it in only one class (…) precisely what I am seeing is to see in what way the student is building his knowledge when I am thinking about a formative assessment, but in the summative assessment what I want to know is if s/he learned the knowledge or not, if he handles the information or not. So, I do not know how in a [summative] assessment I can ask them to build from what s/he knows, then provide them with information that allows them to modify their model and evaluate it.

As shown in the vignette above, even though James acknowledged the role of the GEM cycle as an approach to engaging students in the construction of their knowledge when thinking with models, he lacked enough of a repertoire to assess students' process of generating and modifying their models. By quoting the GEM cycle, James had already enriched his understanding of the foundations of MBT after the OPDC and recognized each of the steps involved in this practice.

Nevertheless, it is likely that the OPDC did not include enough examples that might have enriched this teacher's repertoire about how to reconstruct their current assessment instruments and reshape them based on an MBT approach that included evaluating model construction and modification.

Table 56 presents the results from the R-LPAL for the theoretical dimension of assessment purpose, content and methods. Each column shows ISTs' level of proficiency before and after the OPDC.

Table 56

Results from the R-LPAL for the Theoretical Dimension of Assessment Purpose, Content and Methods

	S		J		E		L		G	
- Asks students to make judgements in science based on reasoning with their own models										
- Uses assessment to assess the internal consistency or coherence of various models constructed by a student and engage students in a cycle of generation, evaluation and modification of the model										

Note: The letters S, J, E, L and G refers to the first letter of each ISTs' name (Samantha, James, Eliana, Lisa and Gabriel, respectively). Colors: aqua blue = novice; cyan blue = advanced beginner; blue = competent; dark blue = advanced assessor.

Table 56 suggests the level of proficiency observed in ISTs' pedagogy. Based on the analysis of class observation and ISTs' interviews on their practice, I observed that most of the ISTs retained the same levels of proficiency before and after the OPDC, coded as being between novice and advanced beginner for the indicators included in this theoretical dimension related to the purpose of assessment. For example, James in the interview before the OPDC mentioned he often assessed content knowledge of a model using the same variety of summative exams (e g., open-

ended questions) which corresponded to a novice level, whereas after the OPDC, in the interview he enriched his knowledge about how to include and assess the GEM cycle in his pedagogy. In other words, he mentioned the idea of not only assessing content knowledge but also modeling practices. The R-LPAL revealed that Samantha enriched her knowledge of how to use assessment to engage students in modeling. The evidence about how she assessed students' reasoning with a model when being engaged in a GEM cycle is detailed in the second theme which is described in section 4.5.2. She was able to help students elicit their initial models of the periodic table, and revise and modify them. Initially, the R-ASMM revealed that she focused on the use of algorithmic problem solving to judge students' understanding which corresponded to a novice level of proficiency since students were not challenged to make judgments or explanations in science based on reasoning with models. After the OPDC, the analysis of the R-ASMM and interviews suggested that Samantha showed competent levels of proficiency for those indicators referenced above and enhanced her knowledge, for example, about how to assess the internal consistency or coherence of their models. Even though she engaged students in the evaluation and modification of their models of the periodic table, the modification phase was briefly assessed which would have corresponded to an advanced assessor level (see section 4.5.2). In summary, based on the analysis of the R-ASMM, interviews, and R-LPAL, teachers showed limited types of strategies related to the dimension of the purpose of assessment which was related to assessing students' field-specific knowledge. In the case of the chemistry teachers (Samantha and Eliana), judging students' algorithmic problem solving was frequently assessed before the OPDC; however, Samantha decreased considerably this practice after the OPDC and was the only teacher who focused on assessing students' modeling practices.

4.5.1.3 ISTs' Communicate Feedback to Clarify Students' Conceptual Doubts of a Model

This theoretical dimension of communication of feedback refers to the strategies that teachers used to support, enhance, or assist student learning during the assessment. The results suggested that teachers used formative feedback to clarify students' conceptual doubts about a model. Analysis also uncovered that ISTs did not often challenge their students to elicit their models. Instead of promoting the generation of models, teachers provided the correct answer or complemented students' answers with a more elaborated explanation in their feedback. This result might reflect that ISTs were not aware of feedback as a tool to provide relevant information to help students make progress with reasoning with a generated model or enriching their modeling practices. Rather, ISTs used feedback in the assessment process to identify students' understanding of a taught model or disciplinary idea. Surprisingly, the results from the QALMBT-Modeling showed that the total sample of Chilean ISTS frequently communicated feedback in an MBT approach. For example, the average for items 3, 16, 19 and 33 was 4.1 which suggests that ISTs frequently implemented assessment strategies to communicate feedback. Based on these results, it seems that ISTs are not able to differentiate between the role of feedback, for example, in a lecture-based approach in comparison to an MBT approach that requires that teachers guide and mentor their students when reasoning with a model. In other words, ISTs answered that they often provide feedback when they teach a target model even though the analysis of qualitative data suggests that they do not very often engage students in the generation of models. For example, for item 16 ("I communicate the results of the assessment in order to help each student achieve a better understanding of the expected model that I want them to learn."), the average in ISTs' answers was 4.2. In another example that involved the

generation of models, item 33 ("I use assessment to give formative feedback to students about the phenomenon that they modeled.") had an average of 3.84. Regarding the 5 ISTs who participated in the phase of identification and development of ALMBT, Samantha and Lisa reported a high average for this theoretical dimension (4.3 and 4.5, respectively), followed by James (3.8), Gabriel (3.5) and Eliana (2.8).

The analysis of class observations yielded evidence that ISTs used two main assessment strategies to provide feedback to students (see Table 57). A third strategy was identified only in an analysis of the interviews (formulation of consensus models). Table 57 shows the R-ASMM for the theoretical dimension included in Figure 1 related to knowledge of feedback. The R-ASMM summarizes the strategies observed in ISTs' pedagogy when assessing the students summatively and formatively, before and after the OPDC.

Table 57

Frequency of ISTs' Strategies Observed for the Dimension of Knowledge of Feedback

	S		J		E		L		G	
- Clarifies students' conceptual doubts (conceptual/ curricular model) and explains the answer (e.g., explains content). The teacher immediately provides the answer.	3 (1.5)	18 (4.5)	16 (3.2)*	3 (1.5)	27 (5.4)*	16 (5.3)		3 (1.5)	3 (1)	1 (1)
- Provides feedback to the whole class, focusing on the main mistakes or difficulties that students had in the exam, or focusing on providing an explanation with the full answer.	2	3	9 (3)	2 (2)	1(1)				1 (1)	

Note: The letters S, J, E, L and G refers to the first letter of each ISTs' name (Samantha, James, Eliana, Lisa and Gabriel, respectively). Colors: blank space = not observed; pale red = observed in only one class; pink = observed in two classes; red = observed in three classes; and maroon = observed in four classes or more.

From Table 57, it is detected that Lisa and Gabriel were the teachers who were observed less often using assessment for the purpose of providing feedback about students' models. The most

common assessment strategy observed in ISTs' pedagogy involved clarifying students' conceptual doubts and was observed for each IST before and after attending the OPDC. Samantha (3 times before the OPDC in two classes with an average of 1.5 times and 18 times after the OPDC in 4 classes with an average of 4.5 times per class) and Eliana (27 times before the OPDC in 5 classes with an average of 5.4 times per class and 16 times after the OPDC in 3 classes with an average of 5.33 times per class) were the ISTs who most often implemented this strategy when teaching science (before and after the OPDC). James, Lisa, and Gabriel also included this practice, but it was observed only on a few occasions after the OPDC. This strategy related to the clarification of conceptual doubts occurred when a student asked for an explanation or asked a conceptual question, followed by an explanation suggested by the teacher in which he/she provided conceptual information or new ideas related to the curricular model. Evidence involving examples of transcripts are provided in table 58. The examples indicated below detail how teachers clarified doubts when students asked conceptual questions. It is worth noting that in each example, ISTs immediately provided the expected correct answer instead of assessing students' initial ideas to challenge them to evaluate their current knowledge about the curricular model (e.g., models of chemical polarity, model of the Calvin cycle). This result reveals that teachers with a low level of assessment literacy tend to provide immediate answers to students' doubts in order to clarify the curricular model.

Table 58

Examples of Teachers Clarifying Conceptual Doubts

IST	Lesson	Example
S	4 (After OPDC)	S3: Miss, I didn't understand why chlorine doesn't react with fluorine. T: Because if you realize, fluorine is more reactive than chlorine. Therefore, here you will not be able to generate a chemical reaction. It is less reactive.
E	1 (Before OPDC)	S: Is no electrolyte the same as non-polar? T: No. No electrolyte can be polar or it can be non-polar. Basically, no electrolyte, the only thing that tells you is that it does not form ions, right?

The excerpts included in Table 58 correspond to transcript of the class observation. I noted that teachers communicated this type of feedback to their students after formatively assessing students' current knowledge with the purpose of reinforcing the disciplinary content or curricular models studied in class as it is indicated in the underlined portions. The second strategy observed in IST' pedagogy included providing feedback to the whole class, for example, after a summative exam, or formatively during the resolution of classroom exercises. Lisa was the only IST who did not include this strategy as observed in her pedagogy. James was the IST who most often included this strategy, especially before the OPDC (9 times before the OPDC in three classes with an average per class of 3 times, and 2 times after the OPDC). In the case of James, in the third class before the OPDC, he reviewed each of the sections of an exam and explained the correct answer in each item to help the students identify the correct answer during class time. Overall, this strategy occurred less often than the strategy related to clarifies students' conceptual doubts because it was mostly used after implementing summative exams, which only occurred at the end of a unit. The following table shows the question included in the exam by James.

Table 59

Example of Summative Assessment Administered by James

Translation: iii) Analyze the information presented in the following text: The nocturnal moth (*Biston betularia*) is frequently used as an example of the evolutionary process. (…) During the industrial revolution, the trunks of trees, in large cities, were stained black due to the deposit of soot and the disappearance of white lichens, very sensitive to pollution. From then on, dark-colored moths hid better in trees than light ones. Records showed that few light moths survived in cities, while few dark ones survived in rural areas, where there was less contamination. In 1956, the "Clean Air Act" gradually reduced the amount of pollution. After 1956 the number of pale moths increased gradually in urban areas. Explain how this example shows the occurrence of natural selection as a mechanism of evolution.

The following vignette describes teacher's feedback, provided in front of the class and used to clarify what he expected as a correct answer regarding a model of natural selection and changes in the peppered moth population during the industrial revolution. He said:

> T: So, many of you just explained it with your words. You were describing again what appeared in the previous text, without referring to the main elements of natural selection You were repeating what appeared in the text. But you never related the answers to natural selection, to connect it, it [answer] had to refer to the concepts of natural selection.

In the answer guideline that James provided me after the class for this exam, he indicated the keywords that the students must include in their answers to obtain a full score. The following text shows the correct answer suggested by James to score a full answer: "(iii) Because it presents all the elements described by Darwin."; "Variability: moths were not all the same."; "Selection pressure: changes in predator behavior / changes in the environment, in the coloration of the logs."; "Differential reproduction: first the dark moths, and then the light ones reproduced more."; "Heredity: the characteristics (body color) passed from generation to generation." In

addition to providing formative feedback in front of the class (underlined portion indicated above), James also provided written feedback on the summative exam, which is shown in the example below (Table 60).

Table 60

Example of Feedback Given by James After Summative Exam

Example of answer from one of the students.
Translation: In this case, natural selection can be evidenced, since one can see a variation (change of quantities in light and dark moths), ~~gradualism~~ (it occurs gradually not suddenly), reproductive success (~~fertile offspring~~) and natural selection (it can be any individual)

Note: The strikethrough type represents the section of the answer that the teacher indicated as wrong.

Based on the strikeout marks, which were only observed in James' pedagogy, he struck through and underlined the keywords that the students included in their answers in order to highlight mistakes. I hypothesize that he used this strategy to indicate the correct keywords in the answers and help students identify their mistakes. The underline type indicates possible keywords that he used to identify the main elements related to Darwin's theory of natural selection. Similarly, Samantha provided feedback to the whole class after an exam as James did in his pedagogy, and she also included individual feedbacks in the exam. Table 61 shows an example of Samantha's strategy to provide feedback. The cursive text in Table 61 shows the notes that Samantha included to help students identify what she expected as a correct answer and help them realize the elements that were missing in their answers.

Table 61

Example of Feedback Given by Samantha After Summative Exam

Example of answer from one of the students.
Translation: (i) Summarize why the atomic radius decreases over period 3, from sodium to chlorine. Decreases as ionization energy and electronegativity increase. Both are inverse to the atomic radius so if they increase the atomic radius decreases. *Increase the number of protons -> increase the effective nuclear attraction with the same number of shells*.

Note: The cursive type indicates teacher's written feedback to students' answer.

A third strategy related to communicating feedback when assessing students was related to generating a consensus model to clarify and enrich students' ideas when working with a model. This strategy was not observed in ISTs' pedagogy through the implementation of the R-ASMM nor mentioned in the interview before the OPDC, though it was mentioned at least once by each IST after the OPDC. After the OPDC, the five ISTs mentioned in interview the idea of promoting the generation of consensus models as a strategy to share their ideas/models with the rest of the class and provide feedback to the class. For example, Gabriel pointed out the fact that he might engage the whole class in generating a consensus model. He stated:

> (…)[R]egarding the revision of their classmate's models, <u>one way of reaching a consensus for them would be to propose the criteria that these models must include to explain certain phenomena… and eventually, they themselves from the ideas they have be able to observe and the ideas they had</u>, <u>to build the models together</u>.

Similarly, James in the interview after the OPDC reflected on how he provided feedback in a class by asking students to explain diagrams and schemes facilitated by him that represented processes and mechanisms, for example, related to gametogenesis. Before suggesting how he

might use a consensus model in class, during the interview, his comments acknowledged the role of engaging students in the modification of models. He said:

> The way to communicate this formative feedback was only by presenting the ideal or correct model at the end. That is, there was no specific feedback regarding how you have to modify your model, or in what way you have to make your modifications, or what would be the revision that you should make of your model so that you can correct any errors that have been made. It was rather general.

When being asked about how he would assess students' models at the beginning, during and at the end of a unit, James briefly suggested an example of how he might promote a consensus model in his class. He suggested that he might ask students to work in groups before sharing a model. He stated:

> I would think of a model built by everyone that would remain as an initial referential model, and that after a process of inquiry, the students would deconstruct and modify. So, based on this information that was provided, within the group at the group level, we build that model. It could be as a group, or with smaller groups within the class. But the goal is for them to build it up to have it as a reference.

The underlined portion shows that after the OPDC James acknowledged the role that the generation, evaluation ("deconstruct[ion]"), and modification of models has during the process of helping students reasoning with a model. Moreover, it seems that James started to enrich his knowledge and repertoire to assess students. In other words, he became more assessment literate in relation to how to communicate feedback in MBT since he initially mentioned that before

attending the OPDC he used to implement formative feedback only to "[P]resent the ideal or correct model at the end." Similarly, in the interview after the OPDC, when Lisa was asked about how she might engage her class in the generation of a consensus model, she stated she might ask her students to present their models in front of the class while the class contributes and comments on an initial model. She said:

> I would ask the students to come forward and draw their models. If we are talking about photosynthesis and we are talking about chloroplasts (….) they could see and say, ah!, this could have been missing, so if we use this (model), I add what the other model has!

In the case of Samantha, she showed a consensus model after engaging students in the generation of a model of the periodic table. She showed the current periodic table as a final representation to achieve consensus regarding how students organized the element in their models related to the periodic trends. In the interview after the OPDC, she pointed out she presented the current periodic table as a consensus model that was achieved by students in the class. She said:

> There were students who came to the current model on their own. And now, they realize and understand the ordering the elements have and are able to search for information that is there that previously did not mean anything to them. Before it did not make sense to them, today it does, today they know what information is contained there.

The underlined portion above suggests that Samantha used the current model of the periodic table as a tool to help students reach an agreement regarding what was the more accurate model to represent the periodic trends. In other words, based on students' generated models, Samantha used the current model as a formative assessment to help students realize if their models were

organized in the same way and if they were able to explain the same periodic trends based on how they organized the elements. Table 62 summarizes ISTs' levels of proficiency for the theoretical dimension of knowledge of feedback. The results included in Table 62 were based on the results from the R-ASMM, interviews before and after the OPDC, and the revision of ISTs' artifacts (summative exams).

Table 62

Results from the R-LPAL for the Theoretical Dimension of Knowledge of Feedback

	S		J		E		L		G	
- Gives formative feedback to students about their generated models for each of the tasks that involve modeling which support students' achievement of the curricular model										
- Communicates the results of the assessment (e.g., summative) for each task in order to help each student to achieve a better understanding of the expected model										
- Promotes the generation of a consensus explanation that helps students have a similar understanding of the phenomenon under study. The whole class participates in the generation of the explanation.										

Note: The letters S, J, E, L and G refers to the first letter of each ISTs' name (Samantha, James, Eliana, Lisa and Gabriel, respectively). Colors: aqua blue = novice; cyan blue = advanced beginner; blue = competent; dark blue = advanced assessor.

Based on the R-LPAL, no major changes were observed in James', Eliana's, Lisa's, and Gabriel's pedagogy for this dimension in terms of their practice/enactment level. Nevertheless, based on the analysis of their declarative level (answers in the interviews), after the OPDC they mentioned strategies to communicate feedback when engaging students in the generation of a consensus model. Therefore, after the OPDC, these four ISTs started to show an advanced beginner level of proficiency. It is worth mentioning that in the case of Lisa and Gabriel these ISTs showed a novice level of proficiency for the majority of the indicators since they appeared to barely communicate feedback about students' ideas and models as it is reflected on the R-ASMM. Since Samantha was the only teacher who engaged students in the generation, use, and

revision of their generated models (as it will be shown in more detail later in section 4.5.2), she moved from lower levels of assessment literacy to a competent level for the indicators included in the R-LPAL. In conclusion, ISTs used feedback to clarify models taught in class, for example, after a summative exam, but they did not use it to help students refine their understanding of their generated models. Moreover, ISTs often used feedback to provide a more accurate explanation of the target model rather than helping students to realize the missing elements in their explanations.

4.5.1.4 Interpretation of Assessment Allows ISTs to Identify Students' Understanding of a Model

ISTs need to be able to identify if students meet the curricular goals of student understanding. The results detailed below are related to the theoretical dimension of knowledge of assessment interpretation and communication included in Figure 1. Specifically, it refers to the use of the results of formative and summative assessment to gather information about students' understanding. Also, through the interpretation of assessment, teachers can adjust their instruction, for example, through questioning students' reasoning with a model. The current study found that teachers used assessment strategies (e. g., driving questions) to identify students' current level of understanding of disciplinary ideas or models. Moreover, teachers interpreted students' answers and explanations to judge the missing elements in their answers, but the science teachers missed opportunities to anticipate students' alternative ideas and help students identify and enrich their elicited models. Interestingly, after the OPDC, teachers acknowledged the relevance of using students' alternative ideas or initial models, but in their practice (enactment level), they were scarcely observed using students' initial ideas or alternative

ideas to help them progress in their understanding of a model. These findings are consistent with Lin and Chiu's (2010) findings that identified that a chemistry teacher with six years of teaching experience and majored in applied chemistry also struggled to anticipate students' learning impediments and alternative ideas when studying models of acids and bases. In another study, Guy-Gaytán et al. (2019) also found that teachers often missed opportunities to guide students during the elaboration and evaluation of their models. The main results of the analysis of qualitative data from the R-ASMM and interviews are supported by the findings from the QALMBT-Modeling for items 7, 21, 24, and 34 which relate to the theoretical dimension of knowledge of interpretation and communication. For example, for item 7 ("I use results from an assessment to compare how students' ideas about a model have been reshaped." and 21 ("I use assessment to judge students' understanding about the phenomenon to be modeled."), the average for the total sample of Chilean ISTs that participated in the phase of baseline of ALMBT was 3.7 (item 7) and 3.6 (item 21), whereas 3.5 and 2.8 for the 5 ISTs, respectively (being 3 sometimes and 4 frequently). Surprisingly, for items 24 ("I use the results generated from formative assessment to adjust the content of my lessons regarding the model that I expect students to learn.") and 34 ("I use assessment to locate evidence about the missing elements in a model that students have not understood.", the total sample of Chilean ISTs showed an average of 3.9 and 3.9 which reflects that they frequently used these assessment strategies. Interestingly, these results from items 24 and 34 for the total sample were higher than the sample of 5 ISTs (*M* = 3, for each item). Regarding the average of the four items (7, 21, 24, and 34), the self-report from Eliana (1.5) and Gabriel (2.5) was the lowest, whereas James (3.3), Samantha (3.5), and Lisa (4) showed the highest self-report for this theoretical dimension.

Among the 5 ISTs' assessment practices for this theoretical dimension, it was identified that ISTs used questioning to check for students' understanding when reasoning with a model and to identify and activate students' prior knowledge. The R-ASMM as applied to the class observation detected two major strategies that involved driving questions that the science teacher used in their pedagogy that were implemented more consistently before and after the OPDC as noted by the analysis of class observations. From an MBT perspective, driving questions are open-ended inquiry questions formulated by a teacher that intellectually engage students in reasoning with their generated model by challenging them to create models to explain phenomena and revise their models to account for findings (Schwarz et al., 2009). It is worth noting that there are two main distinctions that are necessary to make regarding how ISTs used driving questions to interpret assessment. Firstly, as mentioned in the assessment strategies related to the theoretical dimension of disciplinary knowledge and PCK, ISTs used driving questions to encourage students to generate an initial model. This strategy often occurred at the beginning of a new topic or unit to promote the elicitation of a model or explanation. Secondly, in the case of the theoretical dimension of knowledge of interpretation and communication, ISTs formulated driving questions during a formative or summative activity as a response to students' understanding of a model. In other words, the driving questions were used as a maneuver to formatively assess and explore students' current understanding and reasoning with a model based on the interpretation of an assessment. Table 63 shows the R-ASMM which indicates the frequency of the type of assessment strategies that ISTs used to interpret students' understanding of a model.

Table 63

Frequency of ISTs' Strategies Observed for the Identification of Students' Understanding of a Model

	S		J		E		L		G	
- Formulates a driving question and complement students' answers with a more sophisticated explanation or conceptual/curricular model.	4 (2)	17 (3.4)*	23 (5.8)	3 (1.5)	2 (2)	4 (2)	3 (3)	1 (1)	1 (1)	1 (1)
- Uses driving questions to motivate or judge students' understanding (e.g., rules, conceptual information, nomenclature, key concepts, evaluate prior knowledge, determine percentage).	3 (1)	12 (4)	49 (8.2)**	10 (2.5)	2 (2)		10 (2.5)	4 (2)	4 (2)	1 (1)

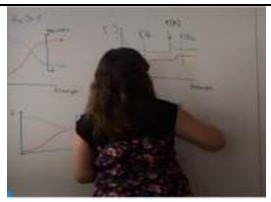
Note: The letters S, J, E, L and G refers to the first letter of each ISTs' name (Samantha, James, Eliana, Lisa and Gabriel, respectively). Colors: blank space = not observed; pale red = observed in only one class; pink = observed in two classes; red = observed in three classes; and maroon = observed in four classes or more.

The first strategy included the formulation of "driving questions" to complement students' ideas about a model or a scientific idea. Samantha was the science teacher who more often used this strategy after the OPDC (17 times in 5 classes with an average of 3.4 times per class). One possible reason for this result is that Samantha valued the importance of engaging students in the evaluation of their models. Therefore, she formulated driving questions to judge students' understanding of a model. This was a common practice that Samantha also used before the OPDC (4 times in 2 classes with an average of 2 times per class). For instance, Table 64 (see "S") shows examples of driving questions that Samantha formulated to ask her students to explain the interaction between molecules to explain gas behavior. Interestingly James used this practice more often before the OPDC (23 times in 4 classes with an average of 5.8 times per class), whereas he used the same strategy only three times in two classes after the OPDC (average of 1.5 times per class). The analysis of the interviews and transcription of the class observation did not provide evidence for the change in James' pedagogy. In the case of Eliana,

Lisa and Gabriel, this action was scarcely observed before and after the OPDC, fluctuating between 1 and 2 times per class. Examples of driving questions implemented by science teachers are indicated in the Table 64. These examples show ISTs' formative assessments of students' prior and current knowledge of a model.

Table 64

Evidence of Driving Questions Formulated by the ISTs During the Class Observation

IST	Lesson	Example	
E	3 (After OPDC)	T: Therefore, if I increase the concentration of reactants, where does the equilibrium shift? S2: Towards the products. T: Towards the products. Therefore, the reactants are going to be consumed again and the products have to increase because more will be formed.	
L	1 (Before OPDC)	T: What is the name of this structure here? They are like stacked discs. The complete structure that is within the chloroplasts. S4: Chlorophyll, right? T: Chlorophyll is the pigment that was inside. But the structure has a name, those that look like discs. Ok, then, can photosynthesis occur in the root? No, why in which part does it occur? S4: In the leaves.	

By questioning students' reasoning, ISTs assessed students' ideas and formulated new driving questions to further explore the coherence of students' understanding of a model. The underlined portion in Table 64 shows examples of questions that ISTs used to help students elicit their models or current understanding. The R-ASMM based on the analysis of classroom observation uncovered a pattern in ISTs' pedagogy, consisting of formulating a question to explore students' current understanding of a phenomenon, process, or mechanism. The teacher then judged students' answers, and finally, complemented the students' initial answer with a more elaborate answer that provided the student with additional content information about the model. These findings are in line with other research findings, for example, Schwarz's (2009), that show that

teachers often focus on the descriptive aspects of a model rather than engaging students in using models as a research tool. The analysis of interviews revealed that ISTs often assessed students' reasoning with models by formulating driving questions before and after the OPDC. For example, Samantha emphasized this strategy in both interviews. In the interview before the OPDC, she indicated she used questions to judge students' learning as indicated in the underlined portion. She said:

> Discussing it, by asking questions, I do not give answers at the beginning because I consider that they are the ones who have to arrive personally to be able to answer them. (…) the idea would be to ask them a question to see that they reinforce the concept that they are answering correctly.

The excerpt above shows that Samantha's knowledge of assessment interpretation was used to reinforce students' knowledge instead of helping students to use and refine their models. In the interview after the OPDC, when Samantha was asked about how she used assessment to involve students in modeling, she mentioned that she used driving questions when engaging students in the evaluation of their models of the elements and periodic trends. For example, during the interview she reflected on her activity in the first class after the OPDC in which she engaged her students in a GEM cycle which reflected a more sophisticated knowledge of assessment interpretation in MBT. She stated:

> These questions were oriented each time to help students try to incorporate new information into the model they had generated and that they themselves realized if the model answered what they were questioning. That is, if at the beginning the elements were ordered according to the valence electrons. Each one ordered it according to their

> criteria. Perfect! Now we go to a data table, take for example the ionization energy, the data that there is for each one of them. Can it [the organization of the elements in the model] correlate with the values of the ionization energy? And there it came a cognitive break, a cognitive challenge. Then, either a different order was generated, or they stayed seated because their model did not fit. (…) Can then another of the properties have a tendency, correlation with the ordering that you have? Therefore, we begin to go further, each time in these questions.

The underlined portion in the vignette above details how Samantha identified students' understanding of their generated models by challenging them to analyze data when evaluating their constructions. By formulating driving questions related to the evaluation of the properties and trends, Samantha was able to interpret students' reasoning with a model and identify the utility of students' generated models.

The second most common strategy that the teachers used included the formulation of driving questions to judge students' understanding of disciplinary core ideas but not necessarily following an MBT approach. Unlike the type of driving questions to explore students' ideas about a model, this second strategy focused on using driving questions to identify students' understanding of a concept (e.g., conceptual information such a definition) that ISTs had already explained in class. In other words, ISTs used this strategy to help students recall information, procedures or concepts. The following table provides examples of the questions that ISTs used to judge students' understanding of disciplinary core ideas.

Table 65

Examples of Driving Questions Used to Judge Students' Understanding

IST	Lesson	Example
J	4 (After OPDC)	What was the name of the following structure, and here I am thinking from the periphery of the seminiferous tubule to the central part of the seminiferous tubule? What were the names that came after the spermatogonia? What did they transform into? S1: Spermatocytes I. T: Right, spermatocyte I.
L	2 (Before OPDC)	So, to synthesize glucose energy is used, how much ATP was used in the Calvin cycle? S2: Two. T: Two, very good. And how many NADPH? S3: One. T: Very good.

Each of the examples indicated above included the formulation of driving questions to reinforce the acquisition of definitions or key ideas studied in class. Table 66 shows the analysis of the level of proficiency of ISTs.

Table 66

Results from the R-LPAL for the Theoretical Dimension of Knowledge of Assessment Interpretation and Communication

	S		J		E		L		G	
- Uses assessment to judge students' understanding about the phenomenon to be modeled and adjusts his/her instruction based on the results obtained from the assessment										
- Uses assessment results to compare how students' ideas about the model studied have been reshaped, for example, within a unit before reaching the reaching the expected curricular model										
- Uses assessment to locate evidence about the missing elements that students have not understood regarding the model under study and adjust his/her pedagogy to coach his/her students during the revision and modification of a model										

Note: The letters S, J, E, L and G refers to the first letter of each ISTs' name (Samantha, James, Eliana, Lisa and Gabriel, respectively). Colors: aqua blue = novice; cyan blue = advanced beginner; blue = competent; dark blue = advanced assessor.

The analysis of the Rubric of Levels of Proficiency (R-ASMM) (Table 66) revealed that ISTs had a limited repertoire to collect and interpret information from assessment when exploring students' models. The R-LPAL shows the changes in ISTs' pedagogy before and after the OPDC for the theoretical dimension of knowledge of interpretation and communication. This table shows that Samantha was the teacher who showed the largest changes in her pedagogy after attending the OPDC for this theoretical dimension related to the interpretation and communication of assessment (from novice to advanced assessor). For instance, after the OPDC it was observed that after reaching a higher level of assessment literacy for this dimension, she used driving questions to locate evidence about how students generated and evaluated their models of the periodic table. Specifically, she used driving questions to help students identify periods, blocks, and groups through the analysis of the electronic configurations of different elements (e. g., "Thinking that the nucleus is going to be here. So, which one has the least energy? The 1s2. Next, which one follows in terms of energy?"). James progressed from a novice level towards an advanced competent level as in his pedagogy he asked his students to generate a model based on the content studied in class. Nevertheless, his approach to teaching was focused mostly on formulating questions to judge students' understanding of key ideas taught in class rather than helping students revise their initial models or explanations. The remaining ISTs remained the same advanced beginner level of assessment literacy (e.g., "Uses assessment to measure students' understanding of the curricular model and inconsistently includes students' ideas from the interpretation of assessment.") In conclusion, the findings showed that teachers had limited strategies to interpret and communicate assessment since they mostly focused on judging students' understanding of ideas previously taught in the science classroom. In this sense, ISTs with low levels of assessment literacy did not use the results of a

formative or summative assessment to identify the missing elements in a model and make decisions about how to adjust their instruction to improve students' reasoning with a model. Rather, they mostly used assessment to explore current understanding of a taught model.

4.5.1.5 Ethics in ISTs' Classroom Assessment Practices Involve the Reinforcement of Students' Elicited Ideas

This theoretical dimension involved ISTs' assessment strategies to give each student the same opportunities to express their ideas and analyze if they understood the models in class. I refer to this theoretical dimension as the fair use of assessment to assess students' reasoning with a model to help students progress in their learning and reflect on their work. As an advanced organizer for this section, the main findings showed that, before and after the OPDC, teachers with low level of assessment literacy rarely offered opportunities to allow their students to express their ideas about a model. ISTs used formative assessment to check students' work when thinking with models; however, this practice was barely observed among IST. When checking students' work, ISTs used this strategy to reinforce ideas studied in class and help them elicit their understanding of a model. The analysis of the results from the QALMBT-Modeling revealed similar results for two of the four items (8, 22, 30, 35) for this theoretical dimension. For example, for item 35 ("I tailor assessment in order to give all students the best opportunities to express their understanding about the model under study."), the average for the total sample of Chilean ISTs who participated in the phase of baseline of ALMBT was 3.4. In the case of the 5 ISTs who participated in the phase of identification and development of ALMBT, the average for the participants was 2.2 which reveals that these group of ISTs very rarely used this strategy in their pedagogy. Another example corresponded to item 30 ("I use the results of the assessment to

coach a student when she/he/they are having problems understanding a model."). On average, the total sample reported an average of 3.75 whereas the group of 5 ISTs reported a mean of 3 (sometimes). While I do not have evidence to support the results for items 8 ("When I develop summative assessment, I inform students in advance about the criteria that I will use to assess their models.") ($M_{\text{Chilean Sample}}$ = 4.15; M_{5ISTs}= 3.6) and 22 ("When students express their claims in front of the classroom, I establish classroom norms to promote a safe expression of students' ideas about their models.") ($M_{\text{Chilean Sample}}$= 4.2; $M_{5\text{ISTs}}$= 4.6) since the majority of teachers did not ask their students to express their models in front of the class, I observed that the 5 ISTs before and after the OPDC promoted a classroom climate of respect which might explain the high ISTs' self-report for these two items. Table 67 shows the R-ASMM for the theoretical dimension of knowledge of ethics of assessment.

Table 67

Frequency of ISTs' Strategies Observed for the Theoretical Dimension of Knowledge of Assessment Ethics

	S		J		E		L		G	
- Checks students' work to identify if some of them are having difficulties thinking with their models. For example, the teacher moves around the class and asks questions.		2 (2)	2 (1)			10 (3.3)	5 (2.5)		4 (2)	
- Uses student's answers to reinforce/reject a conceptual model/prior ideas about a model	1 (1)	4 (4)	5 (1.7)	2 (1)	4 (4)		1 (1)	1 (1)	3 (1.5)	3 (1.5)

Note: The letters S, J, E, L and G refers to the first letter of each ISTs' name (Samantha, James, Eliana, Lisa and Gabriel, respectively). Colors: blank space = not observed; pale red = observed in only one class; pink = observed in two classes; red = observed in three classes; and maroon = observed in four classes or more.

Assessment practices related to this dimension were observed only on a few occasions among ISTs and particularly when they asked students to work in a formative activity as a preparation for a summative exam (e. g., during the resolution of a handout that summarizes the main

contents studied in class). Table 67 shows that two strategies were observed among ISTs to give students equal opportunities to express and clarify their understanding. These strategies included i) checking students' work to identify difficulties when thinking with a model and ii) using students' answers to reinforce or reject their knowledge about a model.

In the first strategy, teachers asked their students to think with a model and checked their work. This assessment strategy occurred when students were reviewing a curricular model. I observed ISTs use the strategy to provide new instructions for a specific activity. Nevertheless, this strategy was barely observed in ISTs' pedagogy before and after the OPDC. In the case of James, Lisa, and Gabriel the strategy was observed only before the OPDC (in two classes) with an average of 2 times per class. Interestingly in the case of Eliana, this strategy was identified only after the OPDC and occurred 10 times in three classes with an average of 3.3 times per class. I did not have evidence that might explain this difference in frequency; however, one possible explanation is that Eliana wanted to prepare students for the summative exam that occurred after the last class observation. Therefore, while students were solving exercises about chemical equilibrium, she checked students' work. In the case of Samantha, she only used this strategy after the OPDC and occurred two times in one class. For example, in the first class after the OPDC, Samantha used this strategy to provide new instructions when the students analyzed information about the electron configuration of the elements. She provided an assorted list of elements from the periodic table and asked her students to identify and highlight the valence electrons. She approached her students before they even generated an initial model of the organization of the elements. She then provided instructions to help students focus on the valence electrons located in the outer shell electron. She stated, "Helium has two, 1s2. Paint the

two electrons on it, please. The neon is going to be $2p^2 2p^6$. Add the two together with the six and paint the eight electrons around the neon." The underlined portion shows the instructions that Samantha provided to guide the student before the generation of their models of the periodic table. Likewise, Eliana was another teacher who checked students' work when they were thinking with a model. Examples of phrases included "Did you finish?", "What will happen then if it increases…. or decreases…?", "Where does the balance begin?", "What happens if I add…?" She used these phrases to guide students individually while they solved exercises for the exam preparation.

Even though ISTs were scarcely observed checking students' work before the OPDC, in the interview before the OPDC, ISTs mentioned this strategy when students worked with models. For example, in this interview when Eliana was asked the question "When students work with models, what is your role in class and how do students interact with you and each other while using models?", Eliana stated she used students' mistakes when solving exercises as an opportunity to monitor her students and clarify doubts. She mentioned:

> (…) [M]y role is usually more of a mediator, in that sense… and what I do is that I go walking, looking, trying to guide a bit and solving doubts (…) seeing at all times what they are working on, what doubts they may have, what erroneous preconceptions they may have. (…) I usually ask them questions, as well, I randomly pick them up and ask them to explain. (…) I don't give them as many opportunities as I should.

From the excerpt indicated above, even though Eliana pointed out that she guides students and helps them resolve doubts, during the class observation before and after the OPDC I observed

that she did not challenge students to revise their initial ideas. I hypothesize that this limited participation when monitoring her students might be because Eliana was not aware of modeling practices related to the revision and testing of models before and after the OPDC. Likewise, before the OPDC, James did not know about his role in giving students opportunities to revise their initial thoughts and models in an MBT approach, as suggested in his interview. In the interview before the OPDC, as it is indicated in the following underlined portion, he mentioned he rarely checked if students were working individually on an activity. He stated, "In general, … I try to keep my intervention as little as possible when students are working. It seems to me that it interrupts them more than it helps them." It is worth noting that I coded the examples indicated above in this theoretical dimension of knowledge of assessment ethics since Eliana and James missed opportunities during a formative assessment to give each student the same opportunity to express and enrich their models when checking students' work. By encouraging students to generate and revise their models, I point out that teachers can enrich the learning environment and promote students' self-confidence when reasoning with models, such as helping students understand the limitation and utility of their models as an opportunity to test, revise and modify their initial ideas or models.

In the interview before the OPDC, when Samantha was asked about how she gave students opportunities to express their understanding of the model that she wanted them to learn at the beginning, during, and at the end of a unit, she mentioned she used driving questions when checking students' work to give each student the same opportunity to express his/her ideas. She said, "Because there is always going to be a student who is good and tends to answer first, therefore, there I have to go measuring, and formulating driving questions to those students that I

see they are behind." This quote (underlined portion) shows how Samantha used formative assessment (driving questions) with the purpose of encouraging each student to express their ideas and have the same opportunities to participate in class. Samantha also mentioned in her reflection in the interview after attending the OPDC that by engaging students in modeling practices she could better understand students' work inside the classroom. She remarked:

> It was clear to me who was in the classroom participating and who was a spectator. Because the truth was that I had no idea. So, that also highlights that one as a teacher already knows who is having difficulties and (…) it gives me more tools to intervene.

The underlined portion above shows that when teachers have higher levels of assessment literacy for this dimension, as it was observed in Samantha's pedagogy, ISTs can identify students' participation and challenge them to express their ideas during the construction of their models. When Samantha was asked about mentioning how she gave her students opportunities to express their understanding of the model under study, she said "Those students who were done and ready to move forward, I gave them the next instruction, and for the remaining students I generated an intermediate question so that they could order or reorganize the data in their generated models [of the periodic table]." I hypothesize that the statement above shows evidence of how Samantha's knowledge of the role of modeling in the science classroom was enriched. She understood that by checking students generated and revised models, she could monitor students' work and understanding of a model in the class. Note that this strategy of checking students' work was observed among all ISTs, but the remaining ISTs did not implement it in their pedagogy based on the R-ASMM and it was only identified during the analysis of their declarative level from the interviews.

Another strategy that was identified through the R-ASMM included reinforcing students' answers in class about a model. This assessment strategy occurred when a student provided a correct definition or explanation related to a model that was complemented by the teacher. This strategy was observed for almost every IST before and after the OPDC. The pattern for this strategy usually included; i) the teacher formulates a question to explore students' understanding about a model, ii) a student briefly provides an explanation or an answer (correct or incorrect), and finally, iii) the teacher reinforces or reject students (e. g., Yes!, Correct!, No!) and provides a more rich explanation. For example, in the following example, in the fifth class before the OPDC, Eliana asked her students to analyze the acid ionization constant of three acids and compare them (HCl, H_2SO_4, CH_3COOH). They said:

> T: Look at the sulfuric acid, which is the first one on top, it dissociates once, right? And then it continues to dissociate, again. Therefore, there are two sources of protons, H^+. They can be added up, therefore, there will be more amount of reactants. Yes?
>
> S1: The reaction speed will be faster.
>
> T: Exactly, then the second one only has one dissociation.
>
> S2: So, it's going to be slower.
>
> T: It will be slower, but not that slower, because it still dissociates a lot. And in the latter, there is practically no dissociation because the constant is very small (…).

In the excerpt indicated above, Eliana reinforced students' answers by providing a more elaborated explanation of a model of acid dissociation. Even though she implemented this strategy to reinforce students' answers, she did not use it as an opportunity to formulate new driving questions to explore students' reasoning either. Rather she used students' answers mostly

to enrich the explanation and convey more content information. This example shown above suggests ISTs' knowledge of assessment ethics was used to reject or reinforce students' answers, which might have influenced students' equitable participation in class. Table 68 shows the results from the R-LPAL for the dimension of knowledge of assessment ethics.

Table 68

Results from the R-LPAL for the Theoretical Dimension of Knowledge of Assessment Ethics

	S		J		E		L		G	
- Provides feedback for each student in order to ensure that each of them achieves the same understanding of the curricular model. For example, s/he summarizes the main points to ensure that each student achieve a similar understanding of the curricular model.										
- Communicates and uses the results of the assessment (e.g., summative) to help each student to achieve a better understanding of the expected model and design specific activities to help student reshape their ideas before moving to the next core ideas that requires a good understanding of the core idea recently assessed**.**										

Note: The letters S, J, E, L and G refers to the first letter of each ISTs' name (Samantha, James, Eliana, Lisa and Gabriel, respectively). Colors: aqua blue = novice; cyan blue = advanced beginner; blue = competent; dark blue = advanced assessor.

From the table it can be noted that ISTs' ethical conducts or actions that might ensure the expression of students' models in the class equitably were slightly enriched after the OPDC only for Samantha and Eliana. Three ISTs kept their initial advanced beginner levels of proficiency since they offered general guidelines to the class to help students improve their understanding of a curricular model but did not check students' thinking with model individually as it was reported by the R-ASMM after the OPDC. Only Samantha and Eliana showed changes in their pedagogy after attending the OPDC. For example, Eliana progressed for the first indicator from an advanced beginner level to a competent level of proficiency ("Communicates and uses the results of the assessment in order to inform the correct answers and also provides some

guidelines to help students revise their models and understanding.") Even though Eliana explored students' understanding of a model, she did not reach an advanced assessor level because she did not promote the generation, evaluation, and modification of models. Rather, she helped students reshape their understanding of the models taught in class. In the case of Samantha, she transitioned into an advanced assessor level of proficiency for this theoretical dimension because she helped students reshape their initial models after the OPDC. Moreover, she checked students' understanding individually not only in the class that she engaged students in modeling but also in the following classes to allow them to apply their understanding of their generated models of the periodic table. The majority of ISTs did not implement strategies to allow their students collectively and individually to express very often their models and only two out of five ISTs enriched the indicators included in the R-LPAL for the theoretical dimension of knowledge of assessment ethics.

4.5.1.6 ISTs have a Limited Repertoire Related to their Knowledge of Scaffolding and Learning Progression to Support Students' Enrichment of Models and Modeling Practices

This theoretical dimension referred to ISTs' strategies to organize and assess content, objectives, and students' models and modeling practices while the complexity of the model increases. In other words, ISTs can help students enrich their ideas and achieve a more complex understanding of a scientific core idea or model. Overall, the R-ASMM showed that before and after the OPDC ISTs mostly focused their learning goals on assessing students' comprehension, recalling knowledge, and analyzing information about a model or disciplinary core idea. This was observed when ISTs summarized content related to a model instead of incorporating scaffolding activities to help students progress in their understanding of a model. The results

from the QALMBT-Modeling also showed that the total sample of Chilean ISTs for two out of five items (9, 12, 14, 18, 25) included in the theoretical dimension of knowledge of assessment and learning progression showed an average of 3.7. This average was observed in question 14 ("When I assess students, I allow them to refine their models to help them reach different levels of complexity about the phenomenon that they are modeling.") and 18 ("I deconstruct a task or objective from the science curriculum into smaller instructional learning experiences that assess students' progression with their models.") These items reflect that Chilean ISTs do not allow students very often to refine their models and they do not implement specific strategies to assess students' progression with their models. Similar results were observed with the sample of 5 ISTs whose average for these two items was 3. In item 25 when being asked about "I encourage students to use their pre-existing ideas in order to help them to construct an initial model that can be enriched later, the average for the 5 ISTs was 2.4 whereas the average for the total sample of Chilean ISTs was 4. A similar trend was observed for item 9 ("I design scaffolded assignments or tasks that progress in complexity in order to asses students' understanding about the model under study"; $M_{\text{Chilean Sample}}$= 3.8; M_{5ISTs} = 3.0). Interestingly, in item 12 "When I make an attempt for students to understand a model, I organize the content in my lessons following a sequence with considers how student understanding can evolve over a span of time"), the 5 ISTs showed the highest average for this theoretical dimension ($M_{\text{Chileanl Sample}}$= 3.6; M_{5ISTs} = 4.06). Overall, three out of ISTs reported an average lower than 3 for the theoretical dimension (Eliana = 1.4; James = 2.8; Gabriel= 2.8) whereas Samantha (3.8) and Lisa (4.2) reported the highest means. The results from the QALMBT-Modeling for the ISTs, excepting Lisa, were related to the results found in the analysis of the R-ASMM, and interviews which showed that teachers barely included and assessed scaffolding and learning progression in their pedagogy. Assessment strategies related to

scaffolding and learning progression were scarcely identified by the R-ASMM. Table 69 shows the R-ASMM for this theoretical dimension.

Table 69

Frequency of ISTs' Strategies Observed for the Theoretical Dimension of Knowledge of Scaffolding and Learning Progression

	S		J		E		L		G	
- Summarizes the main ideas/components (curricular model) that students studied in (last) class in order to help them understand new content and help students make the connection with a new topic.		2 (1)			1 (1)	1 (1)	3 (1)			1 (1)

Note: The letters S, J, E, L and G refers to the first letter of each ISTs' name (Samantha, James, Eliana, Lisa and Gabriel, respectively). Colors: blank space = not observed; pale red = observed in only one class; pink = observed in two classes; red = observed in three classes; and maroon = observed in four classes or more.

Only one strategy was observed in the class observations that was related to summarizing content information studied in the class. This strategy consisted in summarizing the main ideas or components about a curricular model studied in class before introducing new content. This strategy did not occur frequently in the science classroom among these ISTs. I included this strategy in this theoretical dimension of knowledge of scaffolding and learning progression even though ISTs did not implement scaffolding strategies to assess student's progression of modeling practices. As it was indicated in Figure 1, I defined knowledge of scaffolding and learning progression as a dimension that is comprised of assessment strategies related to two domains: i) progression of the understanding of a target model and ii) progression of modeling practices. Therefore, the strategy related to summarizing content information studied in class corresponded to an unsophisticated formative assessment strategy to help students progress in the understanding of a target model just by merely summarizing important information.

For example, in the second class before the OPDC, Lisa summarized the steps of light-independent reactions of photosynthesis. This event occurred at the beginning of the class. They said:

> T: Let's finish with the summary. This process has about six chemical reactions that starts with the fixation, where the rubisco joins with the carbon dioxide. How much carbon dioxide do you need?
>
> S1: Six.
>
> T: Six, very good. Six molecules of carbon dioxide bind with the rubisco and then go through a reduction process where it allows glucose to be synthesized and finally regeneration. In this stage, glucose is released, and the rubisco returns to re-fix CO_2.

The underlined portion in the excerpt above shows how Lisa summarized the steps related to the process of carbon fixation by the action of the enzyme Rubisco. In other words, Lisa included content scaffolding by merely summarizing key information before introducing new content.

Another strategy related to the theoretical dimension of scaffolding and learning progression that was mentioned in the interviews involved the adjustment of the complexity of the models to be taught based on the curriculum goals. This strategy was not observed through the R-ASMM (practice/enactment level), rather it was mentioned in the interview (declarative level) before and after the OPDC when ISTs were asked whether they adapt or reduce the complexity of the model that they want students to learn when the performance of the students is not as expected. For example, in the interview before the OPDC, Gabriel shared that he does not often reduce the complexity of the content he teaches, but instead he adjusts his pedagogy to help students reach the target disciplinary core idea. He said, "Never reducing complexity. I believe that at least they

can achieve that complexity, only that the path that I have shown them is not the most appropriate." In the same interview, he also explained that he had never implemented this strategy when teaching modeling, but he suggested an example based on the model of photosynthesis. He said "I would divide the teaching of the Calvin cycle, because, of course, understanding the whole process of one is like in an hour and a half, I know it is not adequate. Maybe, splitting it out, taking a little more class." In this vignette, Gabriel acknowledged that on some occasions he might need to adapt the assessment activities when the curricular model has to be taught in a short period of time. Similar ideas were suggested by James who was also a biology teacher. He mentioned in the interview before the OPDC, "No, I do not modify the complexity of the model. What I do is rather, make sure the students are able to make the model as it was originally proposed." He also suggested in this interview the use of different models to represent the same phenomenon and help students understand a complex idea or the components detailed in a model. During the interview before the OPDC he said, "I modify, that is, I move to another similar model, of similar complexity, but with different characteristics, or I use a simpler model." Likewise, Samantha also emphasized in the interview before the OPDC that she often tried new pedagogical resources when teaching models but their complexity remained the same. She said "I do not diminish the complexity. Students have to reach that level of complexity, but in that or in different ways. I adapt the material, my class, the form, or a web page, or use different resources." In other words, Samantha mentioned she adapted her pedagogy, for example, the strategies that she used to teach a model, with the goal of helping students reach the understanding of the target model that she initially intended.

Regarding modeling practices, Lisa also mentioned in the interview before the OPDC that she struggled to engage her students in reaching more complex skills such as analyzing and creating. It is worth noting that one of the more important modeling practices is related to the generation of models; nevertheless, it can be noted that before the OPDC, Lisa probably did not have a sophisticated repertoire to engage students in this activity since she mostly focused on conveying content information about the target model. She mentioned during the interview, in the underlined portion, that the learning goals that she often assesses in class, namely "the learning goals of the national curriculum are like analyze and create, and I cannot do that the students achieve them in a class, so in the end, all my objectives are to identify, to compare, to recognize." In the case of Eliana, she was the only IST who mentioned that she selected and adjusted the complexity of the models based on the utility that specific content or curricular models would have had for her students' educational trajectory. She said in the interview before the OPDC:

> For example, when I teach acid and base, we have the theories, right? And you make a model based on the three theories. Normally, the one that it is harder for students is Lewis's, because it has valence electrons, because they don't know Lewis electrons, which is another model. So, what I'm trying to do there is I do not emphasize that part and try to keep them with the other two well-known models, or the other two (…) they are not going to use it [Lewis] that much right now, or until they go to university.

The underlined portion in Eliana's answer suggests that she assessed core ideas or curricular models based on the utility that this new knowledge might have for students in preparation for

later courses and materials. In the interviews after the OPDC, ISTs did not suggest new ideas about how they might adjust curriculum goals to teach models.

The last strategy identified in the interviews was related to the progression of assessment and the organization of the content. By designing scaffolding activities and selecting specific curriculum goals that increase in complexity, ISTs can implement differentiated assessment strategies to assess students' learning progression when thinking with models. Nevertheless, ISTs did not show elaborate answers when being asked about how they assessed how students can enrich their inquiry skills while thinking with models. For example, in the interview before the OPDC, Gabriel mentioned if he had to assess students' skills and learning progress within a unit, he would implement more complex formative assessment once students had reached more basic skills. He said:

> We start from a more basic skill, and through assessments during each class, making the use of more complex skills. First…, with models it would be at the beginning, it would have to be more related to what the student clearly knows, and then include some problems during the unit to elaborate and use a model.

Interestingly, Gabriel mentioned the progression of skills but did not emphasize specific strategies that he might use, and he did not indicate how those skills might progress within a unit. In the case of Samantha, only after the OPDC, was she able to justify her choices to organize the key ideas and the model to be studied. She said;

> I started this modeling activity having completed electron configuration, not having taught any definition of periodic laws. I had not described any historical approximation

of the arrangement of the elements, nothing! And I just threw them into working with something they knew. And, once they realized that there are trends that are born only from the electronic configuration, I went on to deliver the formal content. Because in the end they got themselves to the current model of the periodic table.

As shown in the vignette above in the underlined portion, Samantha mentioned that she oriented students to achieve small steps before teaching the final model. Once students understood that the electronic configuration was information that they could use to give sense to their generated models and once students had generated, evaluated and modified their models based on the identification of periodic trends, Samantha introduced the current version of the periodic table (curricular model) to help students give sense to their generated models. No evidence of the progression of modeling practices as suggested in Schwarz et al. (2012) was identified among ISTs responses in the interviews before and after the OPDC. ISTs' changes in their pedagogy are shown in the R-LPAL indicated in Table 70.

Table 70

Results from the R-LPAL for the Theoretical Dimension of Knowledge of Scaffolding and Learning Progression

	S		J		E		L		G	
- Incorporates scaffolding activities or tasks which progress in complexity in order to assess students' understanding of the model and encourage them to evaluate and modify their models.										
- Adjusts the complexity of the curricular model to facilitate student' understanding of the system under study and leads students in conversations to enrich and refine their ideas about the model that should study according to the provincial science curriculum.										

Note: The letters S, J, E, L and G refers to the first letter of each ISTs' name (Samantha, James, Eliana, Lisa and Gabriel, respectively). Colors: aqua blue = novice; cyan blue = advanced beginner; blue = competent; dark blue = advanced assessor.

The analysis of the R-LPAL for changes of ISTs' proficiency levels for this theoretical dimension of knowledge of scaffolding and learning progression revealed that James, Eliana, Lisa, and Gabriel did not show major changes in their assessment practices. Even though their ideas about how to ask students to express their pre-existing ideas to help students construct an initial model were enriched after the OPDC, the participants rarely provided evidence in their interview responses about how to design and incorporate scaffolding activities to assess student progress. Like in most of the other theoretical dimensions, Samantha was able to enrich her level of proficiency from advanced beginner to competent after the OPDC. This can be illustrated by the examples that Samantha provided during the second interview in which she explained the purpose of the formative assessment that gradually challenged students to use evidence to revise and modify their models. Regarding the remaining teachers, James and Eliana had an advanced beginner level of proficiency beforehand and remained the same level after the OPDC, for example, "Incorporates activities with a similar level of complexity but on some occasions s/he challenges students to reach higher skills (e.g., students elaborate claims, use evidence, use a model." Similarly, Lisa and Gabriel remained at the same levels of proficiency, for example, in the case of scaffolding activities, they showed a novice level since they incorporated activities that always measured the same skills (e.g., rote learning).

To summarize, the evidence collected from the class observations, interviews, the observational rubric (R-ASMM) and the analysis of ISTs' levels of proficiency (R-LPAL) showed that the wide theme related to the purpose of teachers' assessment practices in modeling contained at least six main pre-existing dimensions included in Figure 1. These dimensions included i) disciplinary knowledge and PCK, ii) knowledge of the purpose of assessment, content and

methods, iii) knowledge of feedback, iv) knowledge of assessment interpretation and communication, v) knowledge of assessment ethics, and vi) knowledge of scaffolding and learning progression. The results provided some insight into the purpose of assessment strategies that ISTs used to assess students' models and modeling practices. For the theoretical dimension of disciplinary knowledge and PCK, it was identified based on the QALMBT-Epistemic that ISTs had a good understanding; however, their PCK was not sophisticated since they rarely engaged students in the generation of models. Regarding ISTs' knowledge of the purpose of assessment, ISTs used formative and summative assessment to measure the acquisition of knowledge about a model instead of helping students to make judgments in science based on their reasoning with their models. For the dimension related to knowledge of feedback, ISTs communicated feedback to clarify students' conceptual doubts about a model, but they were not able to implement assessment strategies to individually help students' revise and modify their models. Rather, they focused on helping students reach a better understanding of the curricular model taught by ISTs in class. This result is consistent with Bouwma-Gearhart et al.'s (2009) study which revealed that two teachers struggled to implement strategies to promote the revision of models when teaching nature of matter using a modeling-based curriculum. The interpretation of assessment was another dimension that was identified in ISTs' pedagogy through the R-ASMM. ISTs formulated driving questions to judge students' understanding of a target model; however, they focused on the memorization of a model rather than assessing students' reasoning with a generated model. For the theoretical dimension of knowledge of assessment ethics, the analysis of interviews and the R-ASMM revealed that ISTs reinforced the understanding of students' ideas studied in class, but they provided limited opportunities to express equally their models in the class. Finally, for the theoretical dimension of knowledge of scaffolding and

learning progression, the R-ASMM identified that ISTs did not implement a variety of strategies to help students progress in their learning about a model. Moreover, ISTs did not develop scaffolding activities and assessments to help students progress in the complexity of their models and modeling practices. Interestingly, the analysis of the R-LPAL for each dimension, revealed that Samantha, an experienced science teacher with a postgraduate in science, showed a more significant improvement in her proficiency levels for the majority of the theoretical dimensions related to assessment literacy. A possible explanation for Samantha's level of proficiency may be her adequate and clear understanding of how to help her students progress in their understanding before reaching the expected curricular model, as identified in the second interview, and the observational rubric. Furthermore, Samantha's years of teaching experience allowed her to reshape her repertoire of assessment strategies more readily to offer opportunities to foster students' reasoning with models in comparison to the other teachers with less years of teaching experience. As it was mentioned in the results from the phase of baseline in assessment literacy in MBT (ALMBT), years of experience was a significant predictor of ALMBT that was significantly and positively related to the dependent variable. This result is aligned with Furtak and Heredia's (2014) results, who found that teachers' prior experience, for example, with designing learning progressions can impact their assessment practices and how they implement modeling activities in their lesson plans. Another explanation might be the fact that Samantha during the interview after the OPDC showed evidence that she thoroughly reviewed the modules from the OPDC, since she used in her responses specific vocabulary included in each of them. It is also important to highlight that the other 4 ISTs still enriched their pedagogy regarding the "declarative level" for each of the theoretical dimensions, which was identified in the analysis of the second interview but their change in the declarative level was not as evident as in Samantha's

case. It is worth pointing out that for the "practice/enactment level", changes in ISTs' assessment practices based on the R-ASMM were scarcely observed after the OPDC, especially in the case of Eliana, Lisa, and Gabriel. I acknowledge that because a smaller number of sessions were conducted with these three teachers after the OPDC, there was a short period of time available for ISTs to read the course materials from the OPDC and, these factors might have limited the opportunities to revise and modify their lesson plans which could have impacted upon ISTs' performance during observation. The results of the second theme are presented in the next section and includes the results related to the theoretical dimension of knowledge of peer and self-assessment.

4.5.2 Theme 2: ITSs' Rarely Promote Self and Peer Assessment in their Pedagogy

Another theme that was identified based on the analysis of interviews, class observations and rubrics was the implementation of self and peer assessment of generated models. As it was mentioned previously, when working with models, teachers not only need to guide students in the process of model construction but also engage students in model evaluation and model evolution. In this process of revision of a model, teachers must be able to foster students' capacity to assess their own and their peer ideas during the process of evaluation of a model to help students understand the epistemology of the nature and purpose of models in science. This theme was related to the theoretical dimension suggested by Xu and Brown (2016) related to the engagement of students in assessment. The main findings revealed that the majority of ISTs, before and after the OPDC, did not challenge their students to evaluate their own and peers' generated models and ideas. Furthermore, ISTs appeared to lack the repertoire to intentionally foster students' modeling practices and critical thinking when evaluating their models.

Samantha was the only case that guided her students in the evaluation of their own models, that in turn, influenced how they judged and reasoned with their models. Nevertheless, none of the ISTs were observed or stated that they implemented assessment strategies to promote peer-assessment to assess others' models or modeling practices. Overall, the results for the theoretical dimension of knowledge of peer and self-assessment based on the analysis of classroom observations, interviews and R-ASMM cohere with the findings from the QALMBT-Modeling. For example, the average for the 5 items (items 4, 11, 13, 17, and 29) included in the QALMBT-Modeling related to this theoretical dimension showed that the total sample of Chilean ISTs who participated in the baseline phase of ISTs' ALMBT was 3.64, whereas for the 5 ISTs it was 3. A breakdown of these five items showed that item 4 ("I challenge my students to develop assessment criteria to evaluate the models constructed by their classmates.") had the lowest mean ($M_{Chilean\ Sample}$ = 2.9; M_{5ISTs} = 2.8). This result is aligned with the findings from the phase of identification and development of ISTs' assessment literacy since none of the 5 ISTs implemented this strategy before and after the OPDC. A similar trend was observed for items 13 ("In my classes I ask students to comment on the models created by their classmates.") and 29 ("I challenge my students to show evidence to support their claims about their models.") with means of $M_{Chilean\ Sample}$ = 3.4; M_{5ISTs} = 2.8 and $M_{Chilean\ Sample}$ = 3.8; M_{5ISTs} = 2.8, respectively. It is worth mentioning that Samantha and Lisa reported a score in the QALMBT-Modeling for this theoretical dimension of 4.2 and 4.4, respectively; however, only Samantha showed changes in her instruction after the OPDC transiting form a novice level to an advanced assessor. The remaining teachers kept their initial novice level of proficiency which was supported by their means in the QALMBT-Modeling for the 5 items (James = 2.2; Eliana = 1 and Gabriel = 3.2).

These results from the phase of baseline show that ISTs very rarely engage their students in peer and self- assessment of models.

Before introducing the assessment practices used to engage students in the evaluation of models, I present the learning goal and target model and detail how the process of generation of a model of the periodic table occurred in Samantha's classroom. I will attempt to use Samantha's case to provide an example of the sequence she used to engage students in modeling with the purpose of providing evidence of Samantha's ALMBT. This example shows how Samantha's level of assessment literacy in MBT impacted her pedagogy. I divided the transcript into three main components related to each phase of the assessment of the GEM cycle. I describe a full GEM cycle from Samantha in this section since the evolution of models can require an iterative process of revision and modification until students can reach a full understanding of the final target model. A narrative of the dialogue for each strategy is provided with quotations from the teacher and students interaction. As has already been mentioned in other sections, Samantha explicitly asked her students to generate a model to explain the organization of the elements from the periodic table. This action was only observed after the OPDC. The episode that I describe in full below occurred in the fifth class after the OPDC and shows the sequence observed in Samantha's pedagogy. This represents a full GEM cycle and reflects the most complex and sophisticated example emerging in this study.

Learning Goal and Target Model: Samantha started her class by informing students of the learning goal of the session. On this occasion, it was expected that students "Evaluate the organization of the elements in the periodic table."; "Looking for patterns – the position of an

element in the periodic table allows scientists to make accurate predictions of its physical and chemical properties." The periodic table can be used as an example of a scientific model since, for example, Mendeleev used the periodic table "[T]o predict the properties of as-yet-undiscovered elements, which were later found to be accurate" (Ben-Zvi & Genut, p. 353). This tool has both explanatory and predictive power; hence, it is essential in the study of chemistry. For example, Mendeleev not only predicted the existence of eka-aluminum, eka-boron, and eka-silicon, which were later named as gallium (Ga), scandium (Sc), and germanium (Ge), respectively, but also properties of unknown elements such as oxidation state, oxy-acid formation, atomic volume or metallic character (Stewart, 2019).

Generation: In this example, the teacher began by giving the instructions for the activities. She provided the symbols of the elements from the periodic table and asked her students to write the electron configuration for each element. While she checked that students worked on this activity, she approached the students and clarified the activity. She said, "Look at all you wrote (electronic configuration) for those 26 elements. Only with the electronic configuration, tell me how you could organize them. You have 26 pieces of paper." By asking students to write the electronic configuration and generate an initial ordering for the elements, Samantha asked her students to create an initial model of the periodic table. Then, she monitored the students and asked them to follow the instructions and assessed their work when generating evidence to create their models. For example, she pointed out:

> T: Now, depending on the (electronic) configurations, organize them. You can make groups, with what criteria you are going to organize them.
>
> S4: I have them organized by the sequence in the periodic table.

T: No! I don't want you to look at this (points to the periodic table).

S4: But I already have them organized.

T: No, give me a criterion.

S4: I don't know, in order of energy?

T: Don't look at the energy, just look at the electronic configuration and tell me how you could organize them.

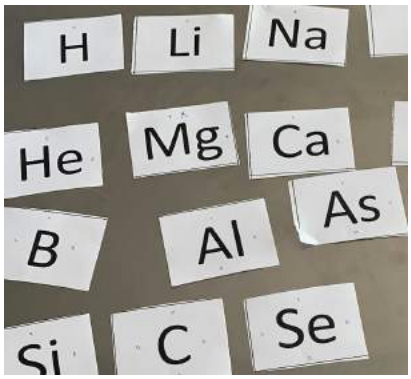
The underlined portion above suggests that Samantha asked her students to create a model by suggesting an initial criterion that helps them justify the order chosen for the elements. At the beginning of the activity, Samantha acknowledged that students might have used other resources to copy the model of the periodic table and organize the elements based on that pre-established organization. She explicitly provided new instructions to clarify the activity. She said:

T: You are going to put the element here, the element that you have, because you are going to take each card and you are going to write down the number of electrons. It has three (electrons) of valence. Then, write Lewis (structure). You are going to paint the electrons as a point around the circle. And, since you are going to have the symbol, you are going to know what element it is. In this case, it is lithium, and now take a picture of what you did so you can compare later.

The underlined portion above shows examples of instruction used by Samantha to guide students in the generation of their models. Figure 21 shows an example of a model generated by a student to organize the elements. It can be noticed that in the first row (H, Li, Na) (Fig. 21. b), the student started to organize the elements based on his/her initial ideas.

Figure 21

Example of Model Generation

a)  b) 

Note: a) Model generated by a student to organize the elements based on their electron configuration. b) Examples of the list of elements that students used to generate their models.

Once students generated their initial models of the organization of the elements, Samantha asked her students to explain their models and justify possible criteria of organization of the elements. They said:

T: Benjamin, tell us what you did.

S1: So, I went from left to right. I did it from left to right and top to bottom. And from top to bottom with the number of electrons they have and how they were arranged.

T: Here you said $1s^1$.

S1: This is the one that have one electron, then 2, 3, 4, 5. (…) Up to nine. Then comes the neon, which is the first noble gas.

T: The neon has $2p^6$.

S1: Which would be until energy level two is completed.

T: Good! You kept going there. And you?

S2: The same.

T: You did the same. How did you end up your model?

S4: In neon.

T: On p^6. And here you started with ... Ah, the $2s^1$ you put it here!. So, here you started in $3s^2$. And here you started in $4s^2$.

S2: No, $3s^1$, $4s^1$.

T: Okay, let's go there. I have two different orders. Because yours is different from his. You ended on p^6, you didn't finish on p^6.

S4: Yes, I did it in order of electrons and when the s was completed here, I started, that is, the first s, then with the 2s in the second level.

In the vignette above it can be noted in the underlined portion that Samantha challenged her students to explain their models in front of the class. She did not judge each model, rather she interpreted the students' organization of the elements. After asking students to generate a model, Samantha promoted the evaluation of students' models of the periodic table. Five assessment strategies were identified mostly in Samantha's pedagogy in the first class after the OPDC and only on a few occasions. These practices were observed after she engaged students in the generation of the model of the periodic table described above. I report these types of practices and their frequency because even though they were rarely observed, they are informative regarding how Samantha's pedagogy changed after the OPDC and how she assessed students' generated models. These five strategies included i) the analysis of data to evaluate a model, ii) the comparison of generated intermediate models, iii) the evaluation of anomalous data, iv) the evaluation of the limitations and utility of a model, and v) the modification of a model. Table 71 shows the frequency and type of the strategies for this dimension based on the analysis of the R-ASMM.

Table 71

Frequency of ISTs' Strategies Observed for the Theoretical Dimension of Knowledge of Peer and Self- Assessment

	S		J		E		L		G	
- Encourages students to evaluate models to identify the utility, scope, and limitations.		4 (2)								
- Asks students to compare their initial models with their intermediate or final models.		2 (2)								
- Ask students to analyze data (e.g., experimental data) in order to evaluate a model and identify, for example, trends, patterns, in order to revise the model. For example, elaborate criteria.		5 (5)								2 (1)
- Encourages students to modify their models, for example, to fit new evidence, new information, or add new elements.		2 (1)								1 (1)
- Engages students in the analysis of anomalous data and help them think with their models or think with a curricular model.		2 (1)								

Note: The letters S, J, E, L and G refers to the first letter of each ISTs' name (Samantha, James, Eliana, Lisa and Gabriel, respectively). Colors: blank space = not observed; pale red = observed in only one class; pink = observed in two classes; red = observed in three classes; and maroon = observed in four classes or more.

The majority of these practices were observed in one or two classes after the OPDC, with a total frequency between 2 and 4 times and an average per class that fluctuated between 1 and 2 times per class. Samantha challenged her students to *evaluate the utility and limitations of their generated models,* for example, through their predictive power. The following excerpt shows an example of a model generated by a student after analyzing the electron configuration. She pointed out:

> T: Do you realize now <u>what would happen if I gave you an element that is a $5p^6$? Where would you go directly to find it in your periodic table?</u>
>
> S1: Here!
>
> T: <u>And in your scheme, would you go find it in your scheme?</u> True? Because you already have your own ordering for you.

The underlined portion above shows how Samantha guided the student to understand the utility of the generated table. Specifically, when she said "[W]hat would happen if I gave you an element that is a $5p^6$? Where would you go directly to find it in your periodic table?", it can be interpreted that Samantha attempted to help the student understand that elements are organized by period and group in which the period corresponds to the principal energy level and the group relates to the extent the subshells are filled. Samantha also provided a new instruction and asked her students to take pictures of their generated models to compare their intermediate models later based on the analysis of the electron energy levels and valency. Students also evaluated the information obtained from the valence electrons. Samantha then helped students think with their models and challenged them to think about the relevance of evaluating their models. She said:

> T: Question for you. If the ordering that you have is reasonable, what more do you ask to that model, that representation? How could you say now, I am going to evaluate, I am going to measure that my model responds to what I want.
>
> S3: When you ask us for an element. How to find it.
>
> T: Very good. And what question should I ask you?
>
> S3: Eh ... which one has a higher ... I don't know, no idea.
>
> T: But I could ask you a question and you could go and find it. What else could you tell me about this model that is useful to me? Because now it is only useful for the electronic configuration. Isn't?
>
> S1: When you ask for more information.
>
> T: When I ask you for more information, right? We are going to see to what extent the model that each one made responds. Ok? Because now I could say to you, with the order you have, I could give you a different property.

In the underlined portions indicated above Samantha explained again why the students needed to be engaged in the assessment of their ideas in order to promote the evaluation of a model ("What else could you tell me about this model that is useful to me?"; "When I ask you for more information, right?) She did not only emphasize the role of the periodic table to organize their elements based on the electron configuration, but also explained that their models of the periodic table must explain the properties of new elements. The underlined segments in the vignette show that Samantha used this formative assessment to help students assess the utility and limitations of their models. To do this, the teacher provided a table with information that included the elements organized by their atomic number (Z). This table also included a set of properties such as atomic radius, ionization energy, electronegativity, electronic affinity, and melting point, which the students used to evaluate their models. In this strategy, she asked her students to *analyze data to evaluate their models*. This episode proceeded as follows:

> T: Look at the table that says ionization. On the sheet. Ok? Now, see how the ordering that you made responds to you. Are they in the same order?
>
> S4: No, they are not in the same order.
>
> T: The model you created yourself is not in the same order?
>
> S4: No.
>
> T: Right. Please look at the electronegativity. It is not ordered, isn't it? Would you like to have everything in order so that these things are also in order and, so, I only learn one model? Another model would have to be built. So, take each card. For each card that you have, look at the number of valence electrons, and that number of valence electrons, turn the card over, identify the element and paint the electrons. Only those of valence electron.

The excerpt above shows that Samantha challenged her students to assess new evidence based on the analysis of atomic properties (e.g., electronegativity) to evaluate their models. It is important to highlight that Samantha formulated questions to help her students reason with their models (e.g., "Are they in the same order?"; "Would you like to have everything in order so that these things are also in order and, so, I only learn one model?") While the students assessed the information, Samantha provided individual and whole class feedback to help students revise their models. She then used formative assessment to ask her students to compare their models and to suggest changes based on the analysis of the table. She said "Whoever has finished, take a look at your ordering. See if you want to change it. And when you're ready, take a second photo of it if you changed it." Samantha promoted new evaluation criteria for their models by asking her students to analyze, in this case, the atomic radius, and challenged her students to *modify their models*. They said:

> T: Take the data sheet. Go to the atomic radius. And now, tell me if for the ordering that you generated, for the number of electrons, the ordering that you put together allows you to show a trend in the atomic radius.
>
> S4: No.
>
> T: It doesn't match. It confuses you. Could you order it in a different way? Do it!
>
> S2: Yes, it does to me!
>
> T: Draw an arrow. To where they increase, or to where they decrease.
>
> S2: It [atomic radius] decreases there [from left to right].
>
> T: Right. Write down, it decreases. And from the bottom to the top, could it be or not?
>
> S3: I don't know.
>
> S4: Do I have to order them to fit the [atomic] radius?

T: Sure. Because now there is no longer a single property, that are electrons. Now, it is with the radius to see if there is a trend.

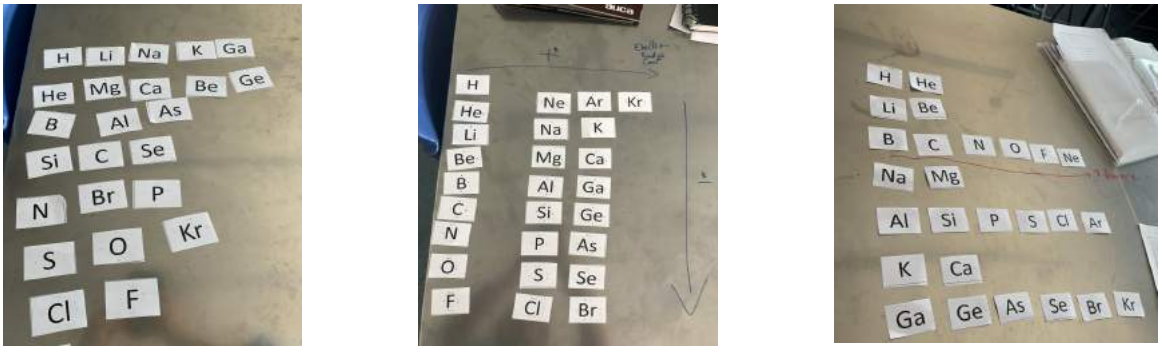
S3: Ok! I do have a trend.

T: You have a tendency. Ok. So, tell me where they are increasing and where they are decreasing. You took a picture of it, right?

This activity concluded the lesson after students had modified their versions of their models. To do so, Samantha asked questions to prompt the modification of the models of the periodic table and challenged her students to return to their models. For example, she said "Tell me if for the ordering that you generated, for the number of electrons, the ordering that you put together allows you to show a trend in the atomic radius." and she asked her students "Could you order it in a different way?" Students' revised models were projected in the next class. Some examples of students' last version of their models are shown in Figure 22. The arrows written by each student show the trends identified by students after modifying their models.

Figure 22

Examples of Different Models Generated by Students



In the second class, the teacher continued conveying content information and clarifying students' conceptual doubts about the model. On some occasions, the teacher referred to students' revised

models to contextualize the class and reinforced students' constructions. For example, at the beginning of the second class after the OPDC, Samantha presented the current version of the periodic table (the curricular model) and pointed out:

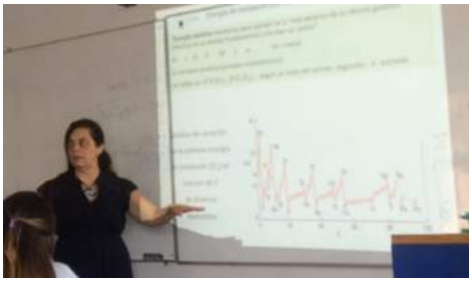
> T: So, you were able to order them by yourself, and one of you ordered it like that and some of you had a different order, but it was fine. That ordering that you created is indicated here (periodic table). So, when I tell you, how many valence electrons does aluminum have?
>
> S1: Three.
>
> T: Which group is it (aluminum) in?
>
> S1: Three.
>
> T: Group three. How many valence electrons does nitrogen have? Group…
>
> S2: 5.
>
> T: Five valence electrons. Or if I told you, tell me the (electronic) configuration of phosphorus. Don't do all this to me. Look there quickly.
>
> S2: Neon.
>
> T: What position is it?
>
> S3: VA.
>
> S1: Ah! It's on level 3.
>
> T: So, I know the electronic differential (last electron) is going to be 3p, and now I fill in forwards.

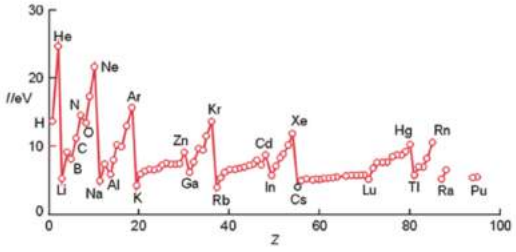
The vignette indicated above shows how Samantha taught their students how to assess the utility and explanatory and predictive power of their models by teaching them how to read the

information provided by groups and periods. For instance, Samantha said "Tell me if for the ordering that you generated, for the number of electrons, the ordering that you put together allows you to show a trend in the atomic radius"; "Which group is it (aluminum) in?"; "How many valence electrons does nitrogen have?" and "What position is it?" After this event, in the same class, Samantha continued conveying content information about the periodic table and explained the utility of the model of periodic table to identify current periodic trends. For example, she provided a diagram of the ionization energy versus atomic number and asked students to generate an initial explanation about the graph (See Figure 23). By analyzing the graph of ionization energy, Samantha asked her students to interpret the periodic table and generate an explanation of the graph.

Figure 23

Graph of Ionization Energy Used by Samantha to Explain Periodic Trends





Note: Consents were provided by ISTs to use their images.

Samantha and her students said:

> T: Look at the graph. Endothermic reactions, I need to take energy to be able to remove that electron from the orbit and ionize the elements. Hence, hydrogen, helium. What could we look for to understand the figure? Look at the figure and tell me how we could begin to interpret this result.
>
> S1: As the lowest with the top is a period. Then like lithium and neon is another period.

T: Very good! So, what could you say about ionization energy as a function of periods?

S1: It is greater to less radius. Isn't it?

T: It is greater when the radius is smaller, therefore, if the radius is smaller, the electrons will be more attracted to the nucleus and it will be harder for you to remove them. Therefore, its ionization energy is higher. Also, that these elements are noble gases and they do not want to lose anything, while those that are below, in which group are they?

S4: In group one.

T: And it's easy for them to release it (electron) and they want to become p^6.

As it can be seen from the example above, Samantha asked her students to analyze the graph and asked them to generate an explanation ("What could we look for to understand the figure? Look at the figure and tell me how we could begin to interpret this result.") Then she reinforced students' answers (S1: As the lowest with the top is a period. Then like lithium and neon is another period. T: Very good!) and formulated driving questions about the curricular model to help students reasoning with the model of the periodic table ("What could you say about ionization energy as a function of periods?") I hypothesize that she used this formative assessment (e. g., asking students to analyze new data and by asking driving questions) to help students reason about the models generated in her first class after the OPDC.

It is also worth mentioning that Samantha was the only teacher as well who suggested *the analysis of anomalous data*. For instance, in the second class after the OPDC, she showed the current version of the periodic table and helped her students think about the limitations of their

model (e. g., “Look where the hydrogen moved, by (electronic) configuration, we had it on the left side. And helium should be in the group above beryllium, but due to the inert behavior of noble gases, it is taken out and put there.”) and re-evaluate their model of the periodic table generated in the first class after the OPDC (e. g., “It is a non-metal [Hydrogen]. And where do we have it located in the (periodic) table? Would you have put it there or moved it?”; "Look where the hydrogen moved, by (electronic) configuration, we had it on the left side.”) Similarly, in the third class after the OPDC where the learning goal was “To analyze the metallic character of the elements in the framework of the periodic table.”, Samantha reviewed the same example, and she stated:

> T: Question for you, what about hydrogen?
>
> S5: It is a non-metal.
>
> T: It is a non-metal. And where do we have it located in the (periodic) table? Would you have put it there or moved it?
>
> S5: I would have moved it.
>
> T: You would have moved it, me too. All them organized. The yellow ones together (metal) and the blue ones (non-metal) together. We would have put it here. So, another question ...
>
> S1: Since it has a valence electron, it is put in group 1.
>
> T: Very good, that is, depending on the electronic configuration. And the same happens with helium, that by (electronic) configuration we would have put helium next to hydrogen, $1s^2$ electronic configuration, but it has a physicochemical behavior related to a noble gas and that is why it is put here. By electronegativity, where would we have placed the hydrogen?

S1: Between oxygen and nitrogen.

Based on the example indicated above, the underlined section of the vignette shows how Samantha directed students' attention to assess anomalous data (e. g., "Question for you, what about hydrogen?"; "It is a non-metal. And where do we have it located in the (periodic) table? Would you have put it there or moved it?") It can be observed that Samantha emphasized that the position assigned to hydrogen in the periodic table can be considered anomalous with her statement that "[it] is, depending on the electronic configuration. And the same happens with helium, that by (electronic) configuration we would have put helium next to hydrogen." Even though hydrogen is not a metal, it resembles alkali metals and some of its properties are similar, such as its capacity to easily form cations. By engaging in formative assessment as above that fostered students' capacity to assess anomalous data (e. g., S1: Since it has a valence electron, it is put in group 1.) and understanding how this information can be used to validate or test their models, after the OPDC Samantha showed that she was able to promote self-assessment of generated models in her pedagogy during the classroom observations which represented a more sophisticated level of assessment literacy in comparison to the remaining ISTs who did not engage their students in the GEM cycle After generating a model, teachers' assessment strategies seemed to be important to guide the student in the evaluation and modification of their own models through a process of self-assessment of their own models as illustrated in her interactions with students.

During the interview after the OPDC, Samantha informed me of how her ideas about the role of evaluating and modifying models were enriched after the OPDC. For example, she explained her

purpose when engaging students in the analysis of data to assess and revise their models. She said:

> Then, they [students] saw different ways of ordering them and once each one made a proposal, they contrasted with data that show properties. So, my ideas changed in the sense that I realized that students can make the discovery and not deliver everything as a speech, as memorizing content, so they can discover, apply, evaluate and generate their own ideas from this. They were able, for example, to modify the order [of the elements] they did initially, but that proposal came from themselves. And if they don't agree, they change it, but there is an argument behind it.

The underlined section of the above excerpt shows that Samantha valued the importance of fostering students' capacity to assess their own models, for example, by analyzing new information (e.g. "They contrasted with data that show properties [metallic character, electronegativity, electron affinity, etc]."). These ideas were not mentioned in the interview before the OPDC. For example, in the interview before the OPDC when Samantha was asked about how she engages students in the assessment of their own and peer models, she said "I have not evaluated models, as a model itself because they [students] have to apply the models" and "But if I had to do it, I have built rubrics for some activity that has a model, but I did not assess how they [students] evaluated the model itself. This text shows that Samantha before attending the OPDC did not have an elaborated knowledge about how to involve students in the assessment of their models. It is worth noting that the R-ASMM, as shown in Table 71, brings to light that Samantha was the only teacher who engaged her students in each of the assessment strategies identified during the class observations. Surprisingly, these strategies were only

identified in the first and second classes after the OPDC, which clearly showed that Samantha's ideas about how to engage students in the assessment of their models were enriched after attending the OPDC. Gabriel also included two of the strategies, but he did not engage students in the generation of models, and he only asked his students to analyze data to evaluate a predefined model facilitated by the teacher which corresponded to lower levels of assessment literacy in MBT. For example, Gabriel mentioned in the interview before the OPDC he engaged students in class discussions to help them comments on others' ideas. He said "in general, they are open discussions within the class, so, I ask a question, and if the answer is good or bad, I ask another question back. If someone believes that the answer is complete, if something is missing." This example suggests that Gabriel formulated questions to engage students in the evaluation of the models studied in class but not about students' generated models. After the OPDC Gabriel mentioned the same strategy related to promoting class discussion, he said "I first raised the models, models that allowed us to explain evolution, and there I asked them, because I showed them in a disorderly way, which model allowed us to better explain evolution." He also mentioned that he might struggle to engage the whole class in peer- assessment, he said "No, individually! I cannot reach that because each one criticizes each one." Moreover, for the majority of the teachers, even though they did not appear to modify their pedagogy during class observations after the OPDC, the results of the interviews showed that they mentioned these strategies as possible new assessment practices that they intend to implement in their future pedagogy. These changes were identified in the interview after the OPDC. In other words, ISTs' practice/enactment level did not show major changes; however, the analysis of their declarative level from the interviews showed that ISTs enriched their knowledge about why is important to engage students in the evaluation of their own and peers' models. For example, in the second

interview, Gabriel acknowledged the role of critique to help students revise and enrich their constructions based on their peers' comments. He said:

> [T]he student could establish comments on how they would improve that model and finally achieve a final activity where students can go back and see what changes they could make to the models according to the comments made by their classmates.

This underlined portion might offer an opportunity for co-assessment in which students collaboratively assess and reflect upon their classmates' models. Based on Gabriel's thoughts, I suggest that assessment literate ISTs must implement assessment strategies in which they promote discussion, negotiation, and the assessment of students' models and modeling practices. In another case, Eliana in the interview before the OPDC mentioned that many of her students are not prepared to receive critiques or comment on their peers' ideas. She said, "I do not feel that they are so mature to discriminate the idea that I like my classmate, or I do not like him, versus s/he did a good job, or s/he did a bad job, as I feel that you have to teach that technique and then they can do it." Based on Eliana's thoughts, the underlined portion shows that a challenge for assessment literate ISTs is to know how to develop and implement strategies that show students the relevance of assessing their peers' ideas in a learning environment and about how to make constructive comments. Likewise, in the interview after the OPDC, Eliana reinforced this idea again and pointed out, "They [students] are super prone to the use of criticism and do not have a good predisposition to criticism and less from their peers." Eliana also mentioned some strategies that she might use in the future to engage her students in assessing their peers' ideas. For example, she mentioned the relevance of teaching her students

how to focus on specific components of a model, for example, related to their explanatory power instead of assessing superficial visual features of the generated model. She pointed out:

> I intend to do it but with pre-established criteria so they do not focus on assessing if they [models] do not have so many colors, or if it is messy. They will notice that if the model explains or relates to central concepts with more peripheral concepts and that this relationship is well made with definitions.

The vignette above depicts Eliana's assessment literacy regarding how to use modeling to engage students in assessment; this literacy was enriched after the OPDC. In this example, Eliana not only focused on visual aspects of a model such as colors and labels, she also emphasized that students should be able to assess the explanatory power of their models through pre-established criteria whereas in the first interview she was not sure about how to engage students in the assessment of their peers' models. Table 72 shows ISTs' levels of proficiency (R-LPAL).

Table 72

Results from the R-LPAL for the Theoretical Dimension of Knowledge of Peer/Self-Assessment

	S		J		E		L		G	
- Challenges students to develop or use assessment criteria to evaluate the models constructed by their classmates in order to encourage others to reflect about epistemic criteria for good models (e.g., regarding the nature and purpose of a model, scope of a model, limitations, etc.).										
- Asks students to analyze new information to promote the evaluation and modification of models that help them collect evidence to show the utility and explanatory and predictive power of their generated models.										

Note: The letters S, J, E, L and G refers to the first letter of each ISTs' name (Samantha, James, Eliana, Lisa and Gabriel, respectively). Colors: aqua blue = novice; cyan blue = advanced beginner; blue = competent; dark blue = advanced assessor.

The analysis of the R-LPAL showed that the majority of IST had a novice level of proficiency before attending the OPDC. The analysis of class observation through the R-ASMM revealed that the majority of ISTs did not engage students in peer and self-assessment. For instance, none of the ISTs asked their students to evaluate their peers' models, and this action was only identified in Samantha's case after the OPDC. Previous studies on MBT conducted by Khan (2011b) have shown that teachers do not follow a systematic approach to MBT and often miss elements related to engaging students in the evaluation and modification of models. Moreover, the generation phase is implemented almost twice as often as the evaluation of models. These results might explain why the dimension of peer and self-assessment was rarely observed among ISTs. The limited knowledge of peer and self-assessment that the group of ISTs had before and after the OPDC are also reinforced by a group of studies that have revealed that science teachers rarely engage students in assessment when teaching with models (Gray & Rogan-Klyve, 2018; Guy-Gaytán et al., 2019; Kawasaki & Sandoval, 2020; Khan, 2011; Lamar et al., 2018; Vo et al., 2019). Even though the analysis of the interviews showed that ISTs acknowledged the role of engaging students in assessing their peer models during the interview, this practice was not observed in the classroom. In other words, their declarative level was enriched but their practice/enactment level was not improved since the R-ASMM did not show any changes. In other words, ISTs with novice levels of proficiency for this theoretical dimension of ALMBT do not possess a vast repertoire to know how to engage students in formatively assessing their peers' ideas. This practice was not observed before or after the OPDC and might be a practice that ISTs rarely used in their instruction. I hypothesize that another reason might be that ISTs lacked knowledge of how to engage students in the process of co-assessment and did not know how to develop students' responsibility and autonomy to allow them to analyze and critically

assess students' models, arguments or ideas. This hypothesis can be supported by the fact that most ISTs, for example, before the OPDC, were not observed challenging their students to test hypotheses or supply evidence to defend their reasoning with models. Assessment literate teachers in MBT must advocate for the development of students' capacity to assess their own and their peers' ideas, for example, i) in terms of the mechanisms, processes, or elements they include in a model, ii) their explanatory and predictive power, or iii) their consistency with evidence.

Regarding the second indicator included in the R-LPAL related to the analysis of information to promote the evaluation and modification of a model, only Samantha showed changes in her pedagogy. The majority of ISTs before and after the OPDC did not challenge their students to analyze new information to promote the evaluation and modification of models that help them collect evidence to show the utility and explanatory and predictive power of a generated model. In the case of Samantha, she started with an advanced beginner level of proficiency because the analysis of the interviews showed that she encouraged her students to collect and analyze information to evaluate their understanding, for example, to determine the Gibbs free energy and evaluate the spontaneity of a process in a galvanic cell. In the interview before the OPDC she said, students "Can begin to infer whether or not some reactions occur due to a trend that can be identified in the periodic table, as in the case of halogens" and "They [students] can evaluate whether a reaction is spontaneous based on the [chemical] potential that is generated." After the OPDC she reached an advanced assessor level of proficiency since she encouraged her students i) to evaluate their generated models of the periodic table based on the analysis of periodic trends; ii) to compare their generated models after analyzing new information; iii) to modify their

models to explain, for example, the trends in the atomic radius; and iv) to analyze anomalous data to evaluate limitations in their generated models. Based on the examples above, it was observed that ISTs reported limited type of practices and the frequency with which they were observed in the classroom was not as frequent as they reported in the QALMBT-Modeling. Finally, the following section presents the last theme identified from the analysis of the data collected for research question 2.

4.5.3 Theme 3: ISTs' Assessment Criteria and Assessment Instruments Measure Students' Knowledge Rather than Assessing their Reasoning with Generated Models

This last theme identified among ISTs' repertoire is related to scoring tools used to assess students' models. This theme was related to the theoretical dimension of knowledge of grading suggested by Xu and Brown (2016) and depicted in Figure 1 in Chapter 2. For example, by informing students about the assessment criteria, assessment literate teachers can communicate the expectations regarding modeling practices or target model that students should achieve during MBT. These criteria provide a guideline that allow students to be aware of their progress and can potentially work as a formative instrument to help them enrich their models. Assessment literate teachers must develop clear assessment instruments and scoring tools to assess students' achievement of the target model and their alternative or inaccurate ideas. The findings from the analysis of class observations, interviews, and ISTs' assessment instruments (e. g., exams) showed that before and after the OPDC ISTs did not develop and implement assessment instruments to assess students' models. In other words, ISTs did not design instruments to assess modeling products and modeling practices, key elements in model-oriented assessment (Namdar & Shen, 2015). Some discrepancies were found between the results from the QALMBT-

Modeling for this theoretical dimension and the analysis of the phase of identification and development of ALMBT. For example, the mean for the four items (items 5, 23, 28, and 32) included in the QALMBT-Modeling for the total sample of Chilean ISTs was 3.9 which meant they frequently implemented assessment strategies related to this theoretical dimension. In the case of the 5 ISTs, the mean was 3.2 which meant they sometimes used these types of strategies. Specifically, items 23 ("I design different scoring tools to evaluate models generated by students.") and 28 ("When I develop distractors in a test, I have in mind the different alternatives or inaccurate ideas that students might have for the model in question.") reported the smallest mean (3 = sometimes) for the 5 ISTs who participated in the phase of identification and development of ALMBT, whereas for the total sample of ISTs the means were 4.0 and 3.8, respectively. The explanation of criteria used to evaluate students' models (item 5) had the largest mean for the 5 ISTs (3.6) and the total sample of Chilean ISTs (4.2). Eliana and Gabriel reported the smallest mean for the 5 items included in the theoretical dimension, 1.3 and 2.8, respectively, whereas Samantha (3.8), Lisa (4), and James (4.3) reported the largest means. One reason that might explain the high self-report from teachers like Lisa and James is that before the OPDC this group of ISTs designed specific instruments to assess students' models that focused on the understanding of a conceptual idea, but they lacked components related to assessing how students can generate, evaluate, and modify their models. This inference can be reinforced by the examples of exam questions detailed later.

It is worth noting that in the section below I still report examples of exams designed and administered by ISTs to provide evidence that they did not possess a vast repertoire related to assessment criteria and scoring tools to assess students' models. Rather, ISTs used traditional test

items that measured students' conceptual understanding of a curricular model studied in class and assessed disciplinary core ideas that did not engage students in reasoning with a model (e. g., resolution of algorithmic problems). For example, the R-ASMM did not identify any strategy among ISTs' assessment practices related to design and implementation of assessment criteria and assessment instruments to grade students' reasoning with a model and model products. Only two assessment practices were identified in class which were related to provide an example of how students might be assessed and explain the goal of a question to help, for example, an item in an exam, to help students understand what the item was asking for. I did not report the frequency and type of these strategies since they were related to a general approach to teaching science that focused, for example, on the resolution of an exercise. Moreover, they were scarcely observed among ISTs (among one or two times in total per IST before and after the OPDC with an average of one time per class). The following excerpt illustrates the case of Eliana during her third class after the OPDC. Eliana designed a handout as a formative activity that students could use to study for the exam. She said:

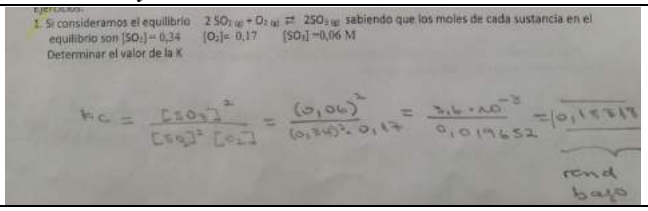
> T: I can ask you something like this. So, I am going to give you a reaction, I am going to tell you that it is exothermic or that its ΔH is less than zero, or I can put a value that is less than zero, and I can ask you what happens to Ke [equilibrium constant] if the temperature decreases.

The underlined portion in this vignette shows that Eliana shared with the students an example of an item that could be included in the exam as a strategy to inform the criteria of assessment. With this example, Eliana informed students that one of the learning goals was that students would be able to numerically evaluate the equilibrium constant (Ke) based on the analysis of

other variables such as temperature and pressure. In class, she said, for example, “evaluate the constant, we put conditions on a system, and you have to answer what happens to them.”, “You have to do the exercise and then make a decision with the value it gives you.” Regarding ISTs’ scoring tools such as summative exams, no changes on this dimension were identified among the assessment instruments before and after the OPDC. For example, in the first class after the OPDC Eliana, administered a worksheet for the topic of chemical equilibrium in which the questions did not involve model-based reasoning since the students mostly evaluated the relationship between variables and interpreted the value of the constant. In other words, for example, students were not asked to interpret chemical equations at the molecular level to explain the constant reshuffle of chemical species. The identification and modification of variables in students’ generated models is often implemented in MBT to help students reason with a model (e. g., Aksit & Wiebe, 2019; Bamberger & Davis, 2013; Campbell et al., 2012; Khan, 2008a; Spier-Dance et al., 2005). In Eliana’s case, the handout started with a short definition about what chemical equilibrium is, and then asked students to write the expression of the constant for different reactions in equilibrium (e.g., $CO_{2(g)} + H_{2(g)} \rightleftarrows CO_{(g)} + H_2O_{(g)}$, determine the value of the constant, and determine if a system was in equilibrium). The following example shows an example of an item included in the formative assessment to prepare her students for the exam.

Table 73

Example of Formative Assessment Administered by Eliana

Examples of answers from one of the students.	
Translation: 1. If we consider the equilibrium $2SO_{2(g)} + O_2 \rightleftarrows 2SO_{3(g)}$ knowing that the moles of each substance are in equilibrium are $[SO_2] = 0.34$; $[O_2] = 0.17$; $[SO_3] = 0.06M$. Determine the value of K Student's answer: $Kc = [SO_3]^2/ [SO_2]^2[O_2] = (0.06)^2/ (0.34)^2\ 0.17 = 3.6 \times 10^{-3}/ 0.019652 = 0.15318$ (low yield)	

In the example indicated above (Table 73), Eliana asked her students to calculate the equilibrium constant and interpret the results. Students were not asked to develop an explanation or a model to explain how a change in the concentration of any product or reactants occurred. In this sense, Eliana's knowledge of grading and scoring techniques used to assess students followed a traditional approach to teaching chemistry that included the resolution of algorithmic problems without asking students to generate and evaluate models of a molecular view of equilibrium. For the exam at the end of the unit after attending the OPDC, Eliana included 25 multiple-choice questions and two open-ended questions to assess students' understanding of chemical equilibrium. This exam covered similar exercises to the one included in the formative assessment for the first class after the OPDC, such as writing and calculating the constant of equilibrium and analyzing graphs. Examples of items included in the exam are detailed in Table 74. The assessment of modeling practices was not explicitly covered in the exam even though she included model products, such as a graph, in which she asked students to identify relationships between variables related to the change of concentration with time. The example included in the assessment appeared to focus on assessing the understanding and interpretation of disciplinary

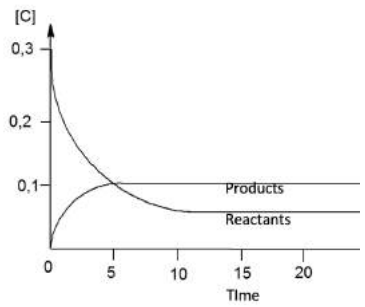
core ideas and algorithmic problem solving rather than assessing how students explained and predicted why the relationships occurred.

Table 74

Examples of Items Included in the Summative Assessment Administered by Eliana

I. Multiple-choice questions.

18. For the reaction of transformation of A ⇄ B the graph of concentration v/s time is as follows;



From the reaction it can be assured that

I) the reaction rate for the reactants is higher at the beginning.

II) at 10 minutes the rate of formation of B is equal to the formation of A again.

III) after 15 minutes the reaction is in a state of equilibrium.

A) only I B) only II C) only III D) I and II E) I and III

II. Open ended questions.

2. Determine the value of the molar concentration of each product in the reaction,

$AgCl_{(s)} \rightleftarrows Ag^{+}_{(aq)} + Cl^{-}_{(aq)}$, if the value of K_{eq} is $4x10^{-8}$ (Consider that the concentrations of Ag^{+} and Cl^{-} are equal.

In the example included in Table 74, Eliana implemented multiple-choice and open-ended questions to assess students' conceptual knowledge and core ideas related to chemical equilibrium. From the example of question 18 above, it can be noted that Eliana asked her students to interpret the graph to assess the understanding of concepts related to chemical equilibrium. I inferred that the type of items included in the exam, as detailed in Table 74, was not cognitively challenging for the students since Eliana often emphasized these ideas and concepts in her classes and asked similar questions during formative assessments to assess students' capacity to recall and interpret information covered in class. In the case of item 2 above as one of the open-ended questions, Eliana also assessed students' algorithmic problem-solving. Before the OPDC, in the summative assessment (exam) at the end of the chemical kinetics unit, Eliana included multiple-choice questions that focused on conceptual problems and algorithmic problem solving. For example, one of the items included in the exam that assessed conceptual

understanding asked "6. The speed of a chemical reaction is much higher when an enzyme is added, compared to when the enzyme is not present. This is because…" whose answers were "A) the enzymes reactivity was altered by the temperature. B) it acts as catalysts that increase activation energy. C) it acts as catalysts reducing the activation energy. D) it demonstrated the formation of products. E) enzymes decrease the reaction energy." This example shows that Eliana assessed the understanding of definitions such as the role of enzymes as biological catalysts that lower activation energy by facilitating bond-breaking through an enzyme/substrate complex. In another example, even within the open-ended questions section, Eliana asked her students to solve an algorithmic problem where students were not challenged to assess models about chemical kinetics (e.g., "Determine the value of the rate constant of the reaction for 2 $NO_{(g)} + 2H_{2\,(g)} \rightarrow N_{2(g)} + H_2O_{(g)}$; [NO] = 0,04 y [$H_2$]= 0.3). In the case of Samantha, her summative assessment instruments also followed a structure comparable to the exams created by Eliana. Samantha included multiple-choice questions that measured algorithmic problem solving, the interpretation of graphs, and the assessment of disciplinary core ideas related to helping students recall definitions. The following example details an item included in one of the summative exams that Samantha administered to assess the unit of stoichiometry and periodic trends. This example occurred in class 4 after the OPDC.

Table 75

Examples of Items Included in the Summative Assessment Administered by Samantha

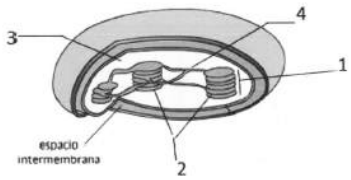
Translation
22. A student determined the percentage of the active ingredient magnesium hydroxide $Mg(OH)_2$, 1.24 g antacid compound. The antacid tablet was added to 50.00 cm^3 of sulfuric acid, H_2SO_4, 0.100 mol dm^{-3}, which was in excess. The tablet completely dissolved. a) Calculate the amount, in mol, of H_2SO_4 added to the sample. b) Write the equation for the reaction knowing that a neutralization reaction occurs, and salt ($MgSO_4$) and water are produced. c) The excess of sulfuric acid was titred, using 20.80cm^3 of standard NaOH solution of concentration 0.1133 mol dm^{-3} so that the equivalence point was reached according to the following reaction: $H_2SO_{4(aq)} + 2\ NaOH_{(aq)} \rightarrow Na_2SO_{4(aq)} + 2\ H_2O_{(l)}$. Find the amount, in moles, of NaOH, and then find the excess, in moles, of acid.

The above example in Table 75 shows that Samantha's summative exam is comparable to that administered by Eliana. In both cases, these chemistry teachers included multiple-choice questions to assess algorithmic problem-solving. In Samantha's example included in Table 75, the open-ended questions were used to assess students' reasoning when solving algorithmic questions. Nevertheless, in the exam, there was no evidence that Samantha was able to include new insight related to modeling based on the modules studied in the OPDC. For example, in the items included in the exam, she included statements like "What quantity", "Calculate the amount", "Write the equation" and "Find the amount of moles" as a strategy to guide students in the calculations; however, she missed opportunities to ask students in the exam to generate, for example, a neutralization reaction model to explain the reactions between an acid and a base and help them understand how these substances interact at the microscopic level rather than merely emphasizing the symbolic level through chemical equations and chemical formula.

In the case of Lisa, the exams that she implemented were predefined by her school. Therefore, she did not participate in the construction of the instruments. Nevertheless, Lisa mentioned in the interviews that she had the possibility of modifying the short quizzes that she often applied administered at the end of some classes. These quizzes assessed disciplinary core ideas and covered specific ideas or concepts about the curricular model that she expected students to learn during the class. These items were multiple-choice and measured the acquisition of knowledge related to the assessment of curricular models. An example of an assessment item designed by Lisa before the OPDC is shown in the following Table 76.

Table 76

Examples of Items Included in the Summative Assessment Administered by Lisa

17. In the following figure, the numbers 1, 2 and 3 indicate respectively: A) Grana – thylakoid – stroma B) Grana - stroma - thylakoid C) Thylakoid - stroma - grana D) Grana - stromal thylakoid- stomata E) Grana- thylakoid - stomata 

In the example detailed above, Lisa asked for specific reactions covered in the class or specific definitions or components in the chloroplast model that students must learn. Even though she included in her exams diagrams to represent the steps involved in photosynthesis, the types of questions included about these diagrams were mostly related to assessing lower order skills such as knowledge (remembering) and comprehension (e. g., identify, recognize, recall). Higher-order questions that involved evaluation (e. g., judge, assess, critique, determine) were largely absent in the exam. In the same way, James also oriented summative assessment towards the understanding and recalling of information related to components or mechanisms of a curricular model. Table 77 shows an example of summative exam developed my James.

Table 77

Examples of Items Included in the Summative Assessment Administered by James

I) Section A. (i) Outline the components of the central nervous system. This can be through: (1) a scheme or drawing, or (2) a text or concept map. Section B . a. Using the information presented in the following table, use your knowledge to identify, circling the concepts that correspond to the four conditions that according to Darwin must be given for Natural Selection to occur. Pressure of selection, Crossing-over, Saltationism, Variability, Inheritance, New genes, Differential reproduction, Mutations.

For example, the learning goals included explicitly for one of James' exams applied before attending the OPDC for the unit of ecology, hormones, reproduction, and nervous system,

specified that students must be able to “i. Explain scientific knowledge”, “ii. Apply scientific knowledge and understanding to solve problems in familiar situations and unknown situations”, and “iii. Analyze and evaluate information to make scientifically-based judgments.” The exams administered by James mostly included open-ended questions. It is worth noting that the summative assessment reviewed below was not administered before the OPDC in 9th grade. From Table 77 it can be noted that James assessed the understanding of a curricular model related to the structure of the nervous system (e. g., outline the components of the central bervous system and make a scheme or drawing). Similarly, in section B of the instrument, students were asked to apply their knowledge of the four conditions in Darwin’s model of natural selection to answer questions concerning how these factors explained the phenomenon of speciation. Even though in the second example students were asked to apply their knowledge about Darwin’s model of data selection, it seems that James focused on assessing the understanding of specific conditions studied prior in class that might have explained the phenomenon observed. It is worth noting that Gabriel did not provide summative assessment instruments that included models.

Having reviewed the evidence of ISTs’ assessment instruments used to grade students and assess their understanding of core ideas and curricular models, ISTs failed to fully assess students’ reasoning with models from an MBT perspective. Assessment literate teachers in MBT must be able to develop diverse assessment instruments and scoring tools to assess how students generate, evaluate, and modify their models at the end of a unit and when students are just starting to study a curricular model. I identified that ISTs mostly assessed their students at the end of the process by applying summative assessments consisting of conceptual questions and

algorithmic questions. ISTs did not implement assessment instruments that capture learning-in-process to identify gaps or misunderstandings when students were thinking with models or when they were enriching their modeling practices. Even though ISTs did not show changes in their strategies during the class observation (practice/enactment level) nor in their developed assessment instruments such as exams, the analysis of the interviews after the OPDC showed that ISTs enriched some of their ideas related to the communication of assessment criteria. Only during the interviews ISTs suggested several ideas about rubrics and criteria to assess their students' models. They also emphasized the role of informing students of the assessment criteria in advance in order to guide them in the process of modeling, reflective of the observation that it was a common practice among teachers before attending the OPDC. For example, in the interview before the OPDC James stated that students must know in advance the assessment criteria when their models and modeling practices are assessed. In the interview after the OPDC, he referred to the same idea and mentioned additionally that he would have to carefully think about the criteria to assess his students' models and inform them in advance. He said "Certain requirements of the model evaluation would have to be met. The instrument can have that guideline, it can have those instructions for the elaboration of the model." Similarly, in the interview before the OPDC, when Gabriel was asked about how he could guide students to reach the expected curricular model, he said "Well first being explicit with the criteria I would expect. It is essential first that they know the criteria with which you are assessing them or what you expect. And, constantly remembering these criteria and using formative assessment." It is worth mentioning again that these ideas were not implemented by ISTs and corresponded to their thoughts about assessing students in MBT. Both ISTs before and after the OPDC, remarked in the interviews that they usually do not assess students' generated models, rather they often

assessed the understanding of final models taught in class as it was identified in the summative assessments developed by ISTs. For example, in the interview before the OPDC, Eliana, said:

> Usually, I don't [assess students' generated models]. But if I did, I would do it with a checklist that is faster, so I can assess everyone… or with some instrument that could be a rubric. I don't think I would assess them through a regular written test. I think the models are a bit difficult to assess.

The underlined portion in the example above lead to the designation that ISTs had a novice level of proficiency in ALMBT, since they rarely implemented scoring techniques and assessment instruments to grade students' models and modeling practices. Regarding their expressed ideas about assessment tools, ISTs suggested some examples in the interviews; for example, Eliana in the interview before the OPDC considered, as an assessment criterion, how students could use a model to solve a problem. She remarked, "[T]he student would have to answer a problem to solve and see if the model is useful for them or not to solve it. With a checklist, then, if it works for them or not." In the case of Lisa, in the interview before the OPDC, she mentioned examples of criteria that she might use to assess students. These criteria focused on assessing elements or components included in a model, for example, to represent the molecule of DNA (see the following underlined portion). She said, "[T]he DNA model have to show the presentation of the structure that presents the DNA and also adds contents, or for example, shows it in a way that is easy to understand." She also mentioned the use of "[A] checklist. I have to build one, what I want the model must have." Specifically, she suggested "[A] checklist, with points. For instance, chromosomes, or condensed genetic material, are indicated. For example, it represents each stage. In the prophase, [s/he] shows that the chromosomes instead of showing decondensed

genetic material." In this example, the underlined portion from Lisa's answer shows that she emphasized the assessment of components and stages to be included in a mechanism or phenomenon as elements she might use to assess students' accuracy of their generated models. The communication of assessment criteria was also emphasized in Lisa's interview after the OPDC She highlighted the importance of sharing the criteria using a rubric (see underlined portion below). She said:

> I wouldn't even show them the rubric at the end, I think halfway through when they start to modify a model, so that they know what they have to include. Because when you show at the end, you realize and they realize, ah, this was what you had to do! So, the rubric gives them a good idea in order to know where they have to go.

In this vignette, the underlined portion shows that Lisa enriched her assessment literacy in MBT by acknowledging her role as a guide in the process of generation, evaluation and modification of students' models. In this interview, she pointed out the role of rubrics to guide students in revising their models (e. g. "I think halfway through when they start to modify a model, so that they know what they have to include.") Samantha also suggested the implementation and communication of a rubric to assess her students when thinking with models. She said:

> I would design and would give students a rubric. Two sections, one that has to do with the content, about concepts, about subject matter, that there is clarity of the definitions, that includes that they can answer all the questions that one is asking them. But the other part, would assess processes in terms of their motivation, compliance, in terms of the rigor of the work they are doing, that they have responded in time.

In this example, Samantha focused on assessing students' alignment of their generated models with the target curricular model (e. g., "content, about concepts, about subject matter, that there is clarity of the definitions, that includes that they can answer all the questions that one is asking them") and valued students' participation and commitment to the activity. During the interview, Samantha used, as an example, her first class after the OPDC, in which she engaged students in the generation of a model to organize the elements based on their periodic trends. In the interview after the OPDC, she mentioned an example of criteria to include in the rubric "[F]or this activity, for example, number of components that match periodic trends, that the model can respond jointly to the interpretation of all the trends of the periodic properties that we are studying." The underlined portion shows examples of two indicators that shows how Samantha started to identify how to define criteria to assess students' models (e. g., number of components that match periodic trends). Even though ISTs did not implement new instruments to assess students' models alone in the science classroom, the results from the interviews after the OPDC showed that ISTs were able to suggest some examples of indicators to be included in their scoring tools for models. Based on the analysis of ISTs' summative assessments and their reflections in the interviews before and after the OPDC, it can be inferred that ISTs were more traditional when assessing their students. They often developed content-based assessments to measure or determine students' learning (for example, recalling and applying). ISTs lacked a sophisticated repertoire not only to engage their students in modeling practices but also to assess them through formative and summative assessments. For instance, in the interview after the OPDC, Eliana acknowledged she struggled to think about how to develop instruments to assess her students during MBT. She said "I have a hard time thinking about how to assess models. I feel that models are super personal for each one, so sometimes I evaluate them in a summative

way, where they have to comply with certain norms." This vignette reflects that Eliana's assessment literacy in MBT and capacity to grade students' models and modeling practices using scoring tools was limited. The underlined portion above shows that she struggled to assess mental models since they are individual constructions. Regarding ISTs' levels of proficiency in ALMBT for this theme related to the communication of assessment criteria to assess students' models, the R-LPAL did not uncover significant changes among ISTs and ISTs had a novice level of proficiency. Table 78 shows the R-LPAL for the theoretical dimension of knowledge of grading.

Table 78

Results from the R-LPAL for the Theoretical Dimension of Knowledge of Grading

	S		J		E		L		G	
- After explaining the criteria, engages students in using the criteria to self-evaluate their own models.										
- Uses different scoring tools (e.g., rubrics, checklists, standards) to evaluate the models generated by students and the development of modeling practices.										

Note: The letters S, J, E, L and G refers to the first letter of each ISTs' name (Samantha, James, Eliana, Lisa and Gabriel, respectively). Colors: aqua blue = novice; cyan blue = advanced beginner; blue = competent; dark blue = advanced assessor.

Taken together, the analysis of class observations, interviews, and the rubrics, showed that IST's levels of proficiency (Table 78) remained at the novice level before and after the OPDC. In other words, ISTs with low levels of assessment literacy struggle to develop and implement different scoring tools (e. g., rubrics, exams) to assess students modeling practices and model products. Even though the analysis of ISTs' declarative level from the interviews showed some enrichment for the theoretical dimension of knowledge of grading, there was no evidence of how ISTs redesigned or developed new instruments to grade students' reasoning with models. Moreover, the majority of ISTs did not engage students in self-evaluation of their generated model and only

Samantha involved students in the GEM cycle as was mentioned in the second theme related to how ISTs promoted self and peer assessment in their pedagogy. Hence, Samantha in the interview after the OPDC was able to briefly suggest an example of criteria to assess students models of the periodic table (e .g., "[F]or this activity, for example, number of components that match periodic trends, that the model can respond jointly to the interpretation of all the trends of the periodic properties that we are studying.") which showed that she started to transition into an advanced beginner level of proficiency and might be able in the future to design scoring tools to assess students' models. Regarding Samantha's case, this difference in her pedagogy might be explained by the fact that she had vast experience teaching science and her pedagogical content knowledge regarding how to use models in the classroom was likely more diverse which made it easier for her to suggest assessment criteria during the interview without overthinking the answer. Another factor that might explain the results for the majority of ISTs is that they were not engaged in revising their artifacts administered to assess students during the OPDC. In the OPDC, ISTs were asked to read the course materials to understanding the foundations of MBT, but there was no major emphasis on revising and adapting their assessment artifacts. Also, I could not collect any evidence of ISTs' instruments that followed an MBT approach (e. g., rubrics) because ISTs did not modify their artifacts after attending the OPDC. Nevertheless, some assessment criteria emerged from the analysis of the interviews after the OPDC as it was suggested by Samantha.

To summarize, based on the data analysis from class observations, interviews, and IST's artifacts, this theme revealed that after the OPDC, the majority of the ISTs continued using models to detail ideas that were previously studied in the class and to help students recall specific

components, elements, or mechanisms. In this sense, their low level of assessment literacy influenced the frequency and type of assessment instruments constructed to assess students' models in the classroom. Moreover, ISTs did not show a sophisticated repertoire to assess their students in MBT and could not engage their students in modeling practices through the GEM cycle. Samantha was the teacher who showed the most significant changes in her instruction. She transitioned from novice/advanced beginner to competent-advanced assessor for at least one indicator included in the R-LPAL for the majority of the theoretical assessment literacy dimensions that were included in Figure 1. James and Eliana also improved their level of proficiency in ALMBT after the OPDC and started to enact several classroom practices that were related to an advanced beginner/competent level of assessment literacy, for example, for the dimensions related to knowledge of the purpose of assessment, knowledge of feedback, and knowledge of interpretation of assessment. For the remaining teachers, it seems that they started transitioning from a novice to an advanced beginner level of proficiency over the two or three classroom observations conducted after the OPDC for the dimensions related to knowledge of ethics and knowledge of interpretation and communication of assessment. These results might reveal that ISTs' years of experience might be one of the most important predictors to explain their ALMBT since Samantha showed the best performance after the OPDC, followed by Gabriel and Eliana.

Before conducting my study, we did not have a clear definition of assessment literacy in MBT. Moreover, we did not know the variables that might influence ISTs enactment of assessment in terms of type and frequency of assessment strategies when teaching with models. The results of this study for research question 1 related to whether ISTs' knowledge of models and modeling

(independent variable) was related to ISTs' assessment literacy in MBT (dependent variable) showed that when ISTs have better knowledge of models, they tend to assess their students' models and reasoning with models more often than teachers who know less about models. In this sense, now we know that how much teachers know about models influences both the type of assessment strategies they enact and how often they assess students while engaged in MBT and in a generic approach. Moreover, now we know from the results from the Exploratory Factor Analysis for the QALMBT-Modeling and the analysis of ISTs' pedagogy, that being assessment literate in MBT is comprised of three main components that ISTs should master when assessing students' models in modeling practices. Firstly, teachers should know the purpose of their assessment practices to engage students in modeling. Secondly, teachers need to be able to promote self- and peer assessment of generated models as a common practice in the science classroom since the epistemology of models and modeling in science education requires that students generate, evaluate, test, revise and modify their models. Finally, assessment literate science teachers in MBT should implement and communicate different assessment tools to grade students' learning progression for scientific modeling.

The findings of this study might also suggest that the reshaping of ISTs' assessment literacy regarding models is a process that starts first with the enrichment of teachers' ideas about this approach to teaching science. In other words, the reconstruction of ISTs' assessment practices in MBT does not seem to be a process that occurs simultaneously while ISTs' embrace the foundations of MBT. As an initial attempt at delineating the dimensions or theoretical dimensions related to assessment literacy in MBT, the findings in this study revealed that ALMBT is a complex set of knowledge and skills that ISTs need to think about first and then put

into practice. In this sense, I argue that assessment literacy is a pivotal component of in-service science teachers' pedagogy, and that assessment literacy has a direct impact on students' learning. According to Nersessian (2013), "[L]earning happens when they [students] perceive the inadequacies of their intuitive understandings—at least under certain conditions—and construct representations of the scientific concepts for themselves" (p. 395). Scientific models are also used to illustrate, simplify, and represent scientific concepts in order to make the ideas more understandable to learners (Rogers et al., 2000). During the process of construction of meaning for scientific concepts, I point out that teachers need to be assessment literate in MBT to develop, implement and use the results of different assessments to guide students in the construction of their understanding about the real world by testing and refining their generate models. Assessment literate ISTs need to be able to gather information from assessment to reshape their pedagogy to help students systematize the relationships between different concepts, features, mechanisms, or components of a curricular model through the generation, evaluation, and modification of a model. Hence, assessment in model-based teaching and learning can help students to reflect on their mental models and conceptual frameworks, facilitating the understanding of science, an idea which Vosniadou (1994) called "metaconceptual awareness". In the chapter that follows, I present the conclusions and implications of the study.

Chapter 5: Conclusions and Implications

This study was designed to: 1) determine if ISTs' knowledge of models and modeling is related to ISTs' assessment literacy in MBT, and 2) identify in what ways ISTs' assessment literacy about models and modeling influenced their pedagogy. The main findings are from a baseline questionnaire on assessment literacy and observational and interview data on the development of assessment literacy in MBT. A summary of the discussed findings will be provided in this chapter in response to each research question.

5.1 Research Question 1: Is ISTs' Knowledge of Models and Modeling Related to ISTs' Assessment Literacy in MBT?

The goal of this research question was to identify the nature of the relationship between ISTs' knowledge about the nature of models and modeling and their assessment literacy in terms of MBT. To do so, a "QALMBT questionnaire" was developed and administered to identify a baseline of ISTs' assessment literacy in MBT. The questionnaire included 35 items related to teachers' general and MBT-oriented assessment strategies and 20 items related to their knowledge of the nature of models and modeling in science (QALMBT-Epistemic). To address the overarching research question, I used Ordinary Least Square (OLS) regression methods to identify if QALMBT-Epistemic was a predictor of ISTs' assessment literacy in MBT. In the results chapter, I provided details of the regression model. The results of the linear regression showed that if teachers have better knowledge of models as shown in the QALMBT-Modeling, it is related with more frequent perception of formative and summative assessments on models. Specifically, in the case of the Chilean sample, this predictor on their knowledge of models was significantly and positively related to ISTs' assessment literacy in both a general and an MBT

approach to teaching science. In other words, ISTs' knowledge of the nature of models and modeling was positively related to how often they self-reported they implement specific assessment practices in MBT (assessment literacy in MBT) such as using assessment to judge students' understanding about the phenomenon to be modeled. Other predictors such as the number of topics in science education courses learned in their teacher education program (e. g., nature of science, learning progression in science education, strategies to elicit students' ideas) and number of courses taken on assessment in teacher education were also significantly positively related to ISTs' assessment literacy- both when ISTs reported on generic or traditional forms of science teaching and when they reported on an MBT approach to science teaching. In the case of a MBT approach (QALMBT-Modeling), the number of years teaching science also significantly predicted ISTs' assessment literacy. Taken together, these statistical results reveal that it is possible to predict ISTs' assessment literacy in MBT using the predictor QALMBT-Epistemic. The following are the key findings from the Chilean sample:

a) When ISTs learn more about science education topics such as the nature of science, models, and modeling in science teacher education, they are more likely to self-report assessment practices related to modeling.

b) Similarly, when ISTs have more years of experience teaching science, their ALMBT is predicted to increase significantly.

c) When ISTs have taken more courses on assessment while studying in their teacher education program, their ALMBT is also predicted to increase significantly.

In broad terms, teachers who have a better knowledge of the nature of models and modeling in science, and/or have studied more topics in science, and/or have taken more courses on

assessment, and/or have more years of experience teaching science are more likely to engage in specific assessment practices in the science classroom to assess students' reasoning with models. Since higher scores in the QAMBT-Modeling mean that teachers more frequently assess students' models and implement different types of assessment, it can be assumed that teachers who have better knowledge of the nature of models, have more years of teaching experience, and know more about topics in science education and assessment, are more likely to engage students in a wider range of assessment practices and more frequently assess students' models.

In the case of the Canadian sample, the results of the OLS regression suggested that
a) When ISTs have better knowledge of models, it is positively related with higher self-reports of formative and summative assessment on models and in a generic or traditional form of science teaching.
b) When ISTs learn more about science education topics in science teacher education, they are more likely to implement assessment practices related to modeling.

It is worth noting that the predictors related to the number of assessment courses taken in their education program and years of experience for this sample did not improve the model fit therefore, they were not included in the final model. In conclusion, similarly to what was observed with the Chilean sample, when Canadian ISTs learn more about topics in science education, they are more likely to implement assessment practices related to a generic form of science teaching. The same trend was observed in the QALMBT-Modeling in which the QALMBT-Epistemic predicted ISTs' assessment literacy in MBT even though this variable was not significant. Interestingly, the predictor related to the number of topics studied in science

education significantly predicted ISTs' QALMBT-Modeling scores. Thus, it was found that ISTs' knowledge of models and modeling was positively and significantly related to their assessment literacy in MBT. This conclusion was complemented by analyzing 5 Chilean ISTs' class observations and interviews.

It is worth also mentioning that the findings from the questionnaire highlight the need for guidance regarding specific assessment practices that ISTs need to implement more often in their pedagogy. The following are the main findings:

a) In a generic approach to teaching science, ISTs report they assess their students more frequently and implement more varieties of assessment practices in comparison to an MBT approach.

b) The factor revealed in the Exploratory factor analysis (EFA) related to the "Implementation of strategies to promote the elicitation and assessment of students' models", which was comprised of six out of eight dimensions included in Figure 1, showed that ISTs frequently included in their pedagogy many of the assessment practices. For example, regarding ISTs' knowledge of assessment purpose, content and methods, it was identified for item 2 that ISTs reported they frequently align their assessment with goals of the science curriculum when assessing the expected models that students should learn (M_{Canada} = 3.9; M_{Chile}= 4.0). In relation to the knowledge of feedback, the clarification of common students' misconceptions or alternative ideas about their generated models in class after a summative assessment (item 19) seems to be a common practice among ISTs (M_{Canada} = 3.9; M_{Chile}= 4.4). Another practice with high self-report was related to the theoretical dimension of knowledge of assessment ethics (item 22). ISTs reported they often establish classroom norms to promote a safe expression of students' ideas

about their models when students express their claims in front of the classroom (M_{Canada} = 3.9; M_{Chile}= 4.2). Similar trend was observed in question 12, for the theoretical dimension of knowledge of learning progression. ISTs reported they often organize the content in their lessons following a sequence that considers how student understanding can evolve over a span in time when they make an attempt for students to understand a model (M_{Canada} = 3.9; M_{Chile}= 4.1).

c) The results from the questionnaire showed that the factor of "Communication of assessment criteria to assess students' models", which was comprised of three items (5, 7, and 8), included two of the most frequent form of modeling assessment. The majority of ISTs reported i) they frequently explain to students the criteria that they will use to evaluate their models (item 5) (M_{Canada} = 3.9; M_{Chile}= 4.2).and ii) they frequently inform students in advance about the criteria that they will use to assess their models when developing a summative assessment (item 8) (M_{Canada} = 3.9; M_{Chile}= 4.2).

d) Some of the assessment practices that were related to the GEM cycle were the form of modeling assessment that were the least frequently implemented by ISTs. For example, for the dimension of disciplinary knowledge and PCK, specifically, items related to the generation of models such as i) using assessment to measure how students develop models to guide their investigations (item 6) (M_{Canada} = 3.0; M_{Chile}= 3.5) and ii) including laboratory activities that require the construction of models by students (item 20) (M_{Canada} = 3.1; M_{Chile}= 3.4) showed low means. Two practices from the theoretical dimension of knowledge of assessment purpose, content and methods were also reported less often by ISTs. These included assessing how students make judgement in science based on reasoning with a model (item 27) (M_{Canada} = 3.4; M_{Chile}= 3.6) and using assessment to evaluate the internal consistency or coherence of various models constructed by a student (item 31) (M_{Canada} = 3.1; M_{Chile}= 3.5). Another item that showed

low means, particularly in the case of the Canadian sample, was the item 34 related to the theoretical dimension of knowledge of assessment interpretation and communication. The majority of the Canadian ISTs reported they sometimes used the results from an assessment to compare how students' ideas about a model have been reshaped (M_{Canada} = 3.0; M_{Chile}= 3.7). Finally, for the theoretical dimension of knowledge of learning progression and scaffolding, item 14 also showed low means. In other words, ISTs when assessing students, only sometimes allow their students to refine their models to help them reach different levels of complexity about the phenomenon that they are modeling (M_{Canada} = 3.3; M_{Chile}= 3.7).

e) Two out of the four items included in the theoretical dimension of knowledge of grading showed particularly low self-report for the Canadian sample, whereas in the case of Chilean ISTs their means were close to 4 (frequent). Item 28 and item 32 showed that teachers when designing assessment instruments rarely reflect on students' alternative ideas (M_{Canada} = 3.0; M_{Chile}= 3.8) and they rarely think beforehand how they will interpret the results from students' models after an assessment (M_{Canada} = 3.4; M_{Chile}= 3.7), respectively.

f) The items related to the factor revealed from the EFA ("Self and peer assessment of generated models") and related to the theoretical dimension of knowledge of peer and self- assessment included in Figure 1, were the least frequent form of MBT assessment used by ISTs. Specifically, the least frequent assessment strategies included i) challenging students to develop assessment criteria to evaluate the models constructed by their classmates (item 4) (M_{Canada} = 2.6; M_{Chile}= 2.9); ii) asking students to test their hypothesis by using their models (item 11) (M_{Canada} = 2.9; M_{Chile}= 3.6); iii) asking students to comment on the models created by their classmates (item13) (M_{Canada} = 2.9; M_{Chile}= 3.4); and iv) teaching students how to judge the quality of their explanations based on the consistency of their models (item 17) (M_{Canada} = 3.2; M_{Chile}= 3.6).

Thus, the main findings showed that ISTs might need particular support in strategies related to the implementation of assessment strategies that involve the generation and evaluation of models. The evidence from the questionnaire suggested that ISTs hardly engaged their students in reflecting and commenting on their own and peers' ideas and models which is key in the process of reasoning with models. Moreover, since ISTs scarcely challenged their students to refine their models, it also seems that ISTs need to enrich the repertoire regarding how to promote the modification of models in the science classroom.

5.2 Research Question 2: In what ways do ISTs' Assessment Literacy about Models and Modeling Influence their Pedagogy?

The goal of this second research question was to inquire into how ISTs' assessment literacy in MBT influenced their pedagogy. Five ISTs' pedagogy and their reflections on it were explored through class observations and individual semi-structured interviews before and after an OPDC. Code-counts provided information about the frequency of ISTs' assessment practices to identify the type of strategies that these ISTs implemented. By comparing the type and frequency of observable practice, I was able to use this information from a rubric that captured these frequencies as an assumption of how ISTs might have enriched their knowledge of assessment for each of the theoretical dimensions used to define assessment literacy in this study. The rubric of assessment strategies in models and modeling (R-ASMM) included indicators that were used to organize, define, and identify codable actions mentioned in interviews and directly observed in the classroom for each theoretical dimension used to define ALMBT. Similarly, the rubric of levels of proficiency in assessment literacy in MBT (R-LPAL) was used to suggest a level of

sophistication of ISTs' assessment literacy. The R-ASMM and the R-LPAL did not reveal considerable changes in the patterns related to assessment practices observed for each theoretical dimension before and after the OPDC, specifically for the practice/enactment level (ISTs' actions). Nevertheless, Samantha was an IST case that showed a number of observable changes in her pedagogy that became aligned with an MBT approach after the OPDC. Her case showed that assessment literate ISTs in order to engage students in modeling need to i) know the purpose of their assessment strategies and have an adequate PCK; ii) interpret students' answers and explanations of a model; iii) offer opportunities to allow their students to express their ideas about a model; iv) understand how key ideas and components of a model related to others and evolve in terms of their complexity (e. g., intermediate models); v) promote self and peer assessment of models; and vi) develop assessment instruments to explore students' reasoning with generated models. Samantha's more sophisticated level of assessment literacy in MBT revealed how this knowledge influenced her pedagogy. For example, the R-ASMM for the class observations regarding the theoretical dimension of disciplinary knowledge and PCK about MBT, showed that she included new assessment practices such as asking students to analyze data to generate a model and asking students to create a model; practices that were not observed in her class before the OPDC. Surprisingly, the theoretical dimension related to engaging students in assessment, specifically for the sub-theme related to the evaluation of models, revealed that Samantha included each of the assessment practices only after the OPDC (i) encouraging students to evaluate models of the periodic table; ii) asking students to compare their initial models; iii) asking students to analyze data such as electronegativity to evaluate a model and identify patterns in order to revise the model; iv) encouraging students to modify their models to fit new evidence; and v) encouraging students in the analysis of anomalous data regarding the

ordering of elements in the periodic table to help them think with their models). This change might indicate that when ISTs are more literate about models and modeling (e.g., as a result of taking the OPDC), it influences how ISTs teach with models and assess students' generated models. Even though the remaining four ISTs did not appear to show major and observable changes (practice/enactment level) in their repertoire in assessment, their declarative level based on the analysis of the interviews revealed that several ideas about how to assess their students were enriched after the OPDC. For example, a shift in ISTs' views of the use of modeling in their assessment strategies was reported in the interviews after attending the OPDC. Before the OPDC, ISTs reported in the interview that they merely used models to represent or describe a phenomenon that they were interested in assessing; after the OPDC, more sophisticated views of models and modeling were shared including the fact that models can help students not merely describe a phenomenon or target but are also useful research tools to make predictions. It is worth mentioning that after the OPDC the majority of the ISTs continued using models in a generic lecture-oriented teaching approach where they basically presented the expected curricular model which revealed their low level of assessment literacy. Furthermore, in the vast majority of the cases, before and after the OPDC, ISTs did not assess the predictive power of model, rather they focused their assessments on the explanatory power of students' models.

Even though the primary focus of this study was not to identify the impact of the OPDC on ISTs' pedagogy, it is still important to analyze how ISTs, and particularly Samantha, become more assessment literate. Some studies suggest that teachers are reluctant to embrace new pedagogical strategies (Yip, 2001), and they might require approximately 18 months to show significant shifts in their pedagogy in the science classroom (Martin & Hand, 2009). Hence, the real impact

of the OPDC might have been limited since teachers were engaged in MBT for only one 10-hour online professional course within one month. The results of the analysis of the five ISTs' who participated in this study are also aligned with Passmore and Svoboda's (2012) reflections who point out that science teachers' instruction often includes accepted models and explanations even when they have received instruction in MBT. Also, the analysis of ISTs' artifacts showed that teachers are mostly "traditional" when assessing their students' understanding of models. By "traditional", this orientation refers to a typical focus on measuring the acquisition of curricular models instead of challenging students to generate and use a model to explain or predict a phenomenon. These results in the present study on MBT might have been expected in this regard, since several similar studies (Contreras, 2016; Ravanal et al., 2018), show that Chilean ISTs self-report their high level of engagement in constructivist practice, but they are often traditional in terms of this pedagogy. Another reason for the apparent limited impact of the 10-hour OPDC might be that the ISTs did not finish the mandatory activities, read all of the course material nor reflected on their pedagogy based on the activities included in the modules. Even though I included activities that attempted to gather more data (e. g., their answers to open-ended questionnaires), ISTs did not submit their answers and only read the course materials as they mentioned to me in informal conversations after the OPDC. Teachers had limited availability to engage fully in the OPDC because of their workload. Another significant aspect that affected data collection and teachers' availability to adjust their lesson plans and assessment strategies were the social protests that paralyzed the Chilean' educational system. Specifically, local authorities announced school closures for weeks due to the riots in Santiago, Chile (October-November 2019), which altered the regular school calendar. Because of this political context, many schools adjusted the school calendar, and teachers were asked to cover and compress the

content prescribed in the curriculum within four weeks. Despite this, after the OPDC, ISTs included several new ideas about modeling in their pedagogy. For example, Samantha, James, and Lisa explicitly recalled the phases of the GEM cycle during the interview, which clearly showed they acknowledged this approach to teaching science after reviewing the third lesson of the OPDC. Moreover, the case of Samantha reported here illustrated how she was able to engage and assess her students in a full GEM cycle during the first class after the OPDC.

Contrasting the five cases based on the analysis of ISTs' pedagogy, Samantha was the teacher who showed the greatest gains regarding incorporating new practices in her pedagogy. These gains are suggestive that the OPDC might have had an impact on Samantha, as Samantha did not display actions related to MBT to a great degree before the course. The OPDC might have also contributed to ISTs development of ideas on MBT. For example, ISTs revealed changes in their intentions (declarative level) to engage in MBT in the interview after the OPDC, but they scarcely implemented them in their practice (enactment level). It is worth considering that according to the observations and interviews, ISTs implemented MBT in an intuitive way in which they think they engaged students in modeling practices, but actually, they were observed and reported teaching and assessing predefined curricular models that students needed to learn in a more traditional way. This result showed their limited assessment literacy in MBT. The limited use of modeling practices in ISTs' pedagogy has also been explored by Justi and Gilbert (2002b) who identified that teachers often include models in their pedagogy without being aware of the full value of models and modeling in learning about science. Similarly, the findings in my study showed that ISTs often used models to provide an authoritarian way of conveying core scientific ideas and lacked the repertoire of modeling practices that engaged students in the construction of

their own models. Justi and Gilbert (2002b) state that this type of pedagogy "[A]lso means that students would get a strong message that 'scientific knowledge' is 'out there' and cannot be created by them; hardly a welcome lesson for any potential future teachers and/or scientists!" (p. 1287). Hence, ISTs might need more guidelines to implement MBT in their pedagogy (Khan, 2011). This claim can be rationalized by Campbell et al.'s (2012) study that highlighted that teachers' discursive modes such as negotiating, elaborating, and reformulating have a key role in guiding students in the process of creating models through abductive reasoning. In this sense, the ways that ISTs used models in the science classroom might also suggest that the participants had difficulties in reflecting about the nature of modeling-based approaches in their pedagogy (Crawford & Capps, 2014; Furtak et al., 2012b).

Rigid school settings might also have limited ISTs' implementation of assessment practices in MBT. It seems there is a natural resonance between ISTs' capacity to implement new assessment strategies and the flexibility or autonomy that schools give to ISTs to try new pedagogies. Based on the analysis of the interviews, teachers pointed out that their summative assessments are, in some cases, predefined by schools. Also, they pointed out that the national curriculum and their schools was perceived to prioritize lectured-based teaching in order to cover each of the units from the prescribed curriculum even though the current science curriculum attempts to promote inquiry practices such as modeling. This rigid structure might have limited ISTs' creativity to implement new strategies that require an inquiry-based approach involving and assessing models. Even though the new Chilean science curriculum is transitioning from lecture-based instruction to inquiry-practices in the science classroom (MINEDUC, 2019), the ISTs who participated in this study still show traditional lecture-based methods of teaching and assessing

their students. I also must acknowledge that two of the ISTs were novice science teachers. In the case of Lisa, when this study was conducted, she had been teaching for only one year. In this sense, Chichekian et al. (2016) found that first-year science teachers usually struggle to implement inquiry-based instruction. It is likely that more years of teaching experience, as also suggested in the QALMBT, might help her enrich her repertoire about how to teach and assess students when thinking with models. This was evident in the case of Samantha. Her case revealed that her disciplinary knowledge and PCK and likely the year of experience teaching science were important variables that influenced how she designed assessment strategies and interpreted the results of assessment to engage students in modeling. This idea is supported by research on PCK which has shown that years of teaching experience is linked to ISTs' pedagogical practice (Chan & Yung, 2018; Friedrichsen et al., 2008; Grossman, 1990; Kind, 2009; Schneider & Plasman, 2011). As it was revealed in chapter 4, this qualitative result is aligned with the findings from the regression model which showed that when ISTs have more years of experience teaching science, their ALMBT is predicted to increase positively and significantly. Therefore, they are more likely to assess students' models in the science classroom since they might have more experience designing and implementing instruments to assess students.

Thus, from the current dissertation, it has become clear that ISTs need assistance in MBT if they are interested in enacting this pedagogical approach, particularly in regard to how to implement formative and summative assessments when challenging students to think with models. Based on the analysis of the QALMBT questionnaire, this recommendation can also be extended to both the Canadian sample since they showed similar self-reports of their assessment strategies to the

Chilean ISTs who answered the questionnaire. Moreover, even though the Chilean sample has a higher self-report for many of the items included in the questionnaire, the analysis of the 5 ISTs showed that they rarely engaged and assessed students' models and modeling practices and the differences between the baseline for the questionnaire for the Canadian and the Chilean ISTs could be due to the different contexts rather than the fact that Chilean ISTs actually assess their students more often in MBT. The evidence presented in this study also suggested that ISTs are mostly traditional not only when presenting curricular models but also when designing and implementing formative and summative assessments. Therefore, even though ISTs' might reshape their declarative level and embrace an MBT approach, Chilean ISTs still need assistance regarding how to synchronize how they teach models and modeling practices and how they assess students' reasoning with their generated models even after attending the OPDC. Finally, these results lead us to i) explore in the future to what extent factors such as teachers' motivation, time, and school setting, among others, might influence how ISTs' implement MBT in the science classroom and ii) how these variables might influence ISTs' repertoire to assess students when thinking and using models. These variables were not explored in the QALMBT; however, the analysis of quantitative data and qualitative data revealed that ISTs' teaching experience and their knowledge about the nature of models, and how to teach with modeling might be important predictors of assessment literacy in MBT.

In terms of the online professional development course, the findings of this study might also provide evidence about ISTs' capability of implementing assessment strategies in MBT and at the outset, it appears from the QALMBT questionnaire, observations, and interviews, that science teachers in this study do not possess a sophisticated level of assessment literacy in MBT.

Moreover, based on the analysis of the questionnaire, class observations, and teachers' artifacts (e. g., exams), the findings also showed that teachers assessment literacy in a general approach followed a traditional pedagogy in which teachers focused, for example, on assessing the acquisition of content knowledge rather than promoting critical thinking or higher-order thinking skills. Specifically, the main findings are:

a) The most common assessment practices were related to three theoretical dimensions included in Figure 1 and related to the first theme of "Implementation of strategies to promote the elicitation and assessment of students' models". It is worth mentioning that the majority of the assessment practices included by ISTs lacked the main components related to an MBT approach, such as encouraging students to elicit their models, but they are still reported since they are informative regarding the complexity of ISTs' assessment literacy in MBT. Firstly, regarding ISTs' disciplinary knowledge and PCK, the 5 ISTs provided content information in any form about a curricular model to explain, for example, mechanisms such as gametogenesis or photosynthesis instead of asking students to generate a model. This was a practice that was often implemented before and after the OPDC. Secondly, the theoretical dimension of interpretation of assessment showed another practice that was more frequent among ISTs' pedagogy. Specifically, ISTs used driving questions with two purposes; i) to judge students understanding of content information, which was more aligned to a traditional approach of teaching; and ii) to complement students' answers with a more sophisticated explanation, for example, about a curricular model. In this type of assessment ISTs formulated a driving question in order to assess students' current understanding of a model. Thirdly, another strategy that was also often implemented among ISTs was related to the theoretical dimension of knowledge of feedback. Samantha, James and Eliana included in the majority of their classes, before and after the OPDC,

the clarification of students' conceptual doubts about a model and the explanation of an answer. The feedback provided by these ISTs often focused on teaching the expected curricular model instead of helping students refine their understanding of a model by themselves. This result is consistent with Guy-Gaytán et al.'s (2019) study which suggests that science teachers often display *limited* feedback practices and scarcely promote the refinement of students' generated models.

c) The least frequent form of modeling assessment was related to the theoretical dimension of disciplinary knowledge and PCK, knowledge of learning progression and knowledge of peer and self- assessment. Regarding the first theoretical dimension (disciplinary knowledge and PCK), ISTs' rarely asked students to create a model. In fact, this action only occurred once after the OPDC in the case of Samantha and James, and once in the case of Gabriel before the OPDC. Asking students to analyze data in order to generate a model and identify, for example, trends and patterns was another practice from the dimension of disciplinary knowledge and PCK that ISTs scarcely used. Even though ISTs after the OPDC acknowledged in the interview the role of models in science education, only Samantha explicitly asked students to generate models of the periodic table and used them in her pedagogy. This result is compatible with the finding from item 6 in the questionnaire that showed that almost 50% of Canadian ISTs and almost 30% of Chilean ISTs report they sometimes use assessment to measure how students develop models to guides their investigations) ($M_{\text{Canada}} = 3.0$; $M_{\text{Chile}} = 3.5$). Regarding knowledge of learning progression and scaffolding, it was only observed from the observation rubric (R-ASMM) that ISTs on a few occasions (once per class) summarized the main ideas or components related to a curricular model in order to help students understand new content and help them make the connection with a new topic. Moreover, ISTs did not design scaffolded assignments or tasks that

progressed in complexity. The theoretical dimension of knowledge of peer and self-assessment, which was related to the second theme "ISTs rarely promote self-and peer assessment in their pedagogy", was also scarcely observed among ISTs and only Samantha after the OPDC included it in her pedagogy. These practices involved i) encouraging students to evaluate models to identify the utility, scope, and limitations; ii) asking students to compare their initial models with their intermediate or final models; iii) asking students to analyze data in order to evaluate a model and identify patterns in order to revise a model; iv) encouraging students to modify their models to fit new evidence; and v) engaging students in the analysis of anomalous data to help them think with their models. These practices were observed only in the two classes after the OPDC with a frequency that fluctuated from one to two times per class. This change in Samantha's pedagogy and the enrichment of her responses in the interview such as valuing students' capacity to generate a model of the periodic table and use it to predict trends, showed evidence that her insight after attending the OPDC was enriched. Moreover, this evidence showed that when ISTs are more assessment literate about models and modeling, this knowledge in assessment may influence their pedagogy and the strategies they implement in the classroom.

d) For the third theme that referred to "IST's assessment criteria and assessment instruments measure students' knowledge rather than assessing their reasoning with generated models", which was related to disciplinary knowledge of grading in Figure 1, revealed that ISTs did not reshape their instruments to assess students' models and they did not use any specific assessment practice, for example, regarding using different tools to evaluate the models generated by students and the development of modeling practices. In other words, ISTs' artifacts such as exams focused on asking students to identify mechanisms or features in a model and assessing algorithmic problem reasoning in the case of chemistry.

e) In terms of their practice/enactment level (based on observational practices in the science classroom), the R-LPAL showed changes mostly in the case of Samantha. She transitioned from a novice/advanced beginner for the majority of the theoretical dimension into a competent/advanced assessor. Changes were more noticeable for two of the theoretical dimensions included in Figure 1. First of all, her disciplinary knowledge and PCK was enriched since she explicitly engaged students in the generation of a model of the periodic table after the OPDC, which was not observed before the OPDC. Moreover, she mentioned in the interviews the importance of engaging students in model construction. Secondly, the theoretical dimension of knowledge of peer and self-assessment was also enriched, specifically for the indicator related to asking students to analyze new information to promote the evaluation and modification of models that help them collect evidence to show the utility and explanatory and predictive power of their generated models of the periodic table. On the contrary, the remaining ISTs did not show major changes in their pedagogy regarding their practice/enactment level and only changes in the declarative level from the interviews were identified, specifically for their disciplinary knowledge and PCK. Moreover, no changes in the planning and design level (lesson plans and assessment instruments) were identified among ISTs which showed that teachers ISTs had a novice level of proficiency regarding their knowledge of grading.

Finally, the findings of this study from the analysis of the questionnaire and the observation of the 5 ISTs' pedagogy suggest that when ISTs have a more sophisticated disciplinary knowledge and PCK regarding models and models, this knowledge might influence how often, and the variety of assessment strategies that ISTs implement in their pedagogy to assess students. This result was particularly evident in the case of Samantha. Even though the results from the

questionnaire showed that the 5 ISTs had a good knowledge about models (disciplinary knowledge) based on the results of the QALMBT-Modeling, Samantha's PCK, based on the analysis of her pedagogy and interviews, was more sophisticated that the remaining ISTs. As suggested by the results from the questionnaire and Samantha's case, years of experience also appeared to have an important role in how ISTs assessed their students in the classroom and how they reshaped their pedagogy into an MBT approach after attending the OPDC. In other words, it is likely that Samantha based on her experience, had a more sophisticated repertoire of assessment strategies that allowed her to adapt her instruction. This variable might also explain why Lisa and Gabriel showed the least changes in their pedagogy after the OPDC since they had only a few years of teaching experience.

5.3 Limitations

Several limitations can be identified in this study. One limitation is related to the baseline in assessment literacy in MBT. ISTs in both contexts, Canada and Chile, were contacted individually based on public information available. These strategies might limit the generalization of the results in other contexts. Also, the results from both countries can be compared; however, the sample in Canada was smaller than the Chilean sample. In the case of Canada, the sample of more than 40 participants might have decreased the statistical power in finding significant differences (type II error), and type II errors might lead one to draw an incorrect conclusion about some of the predictors (Hawley et al., 2019; Nayak, 2010). In other words, the results of the test might have indicated that QALMBT-Epistemic was not a significant predictor of ISTs' assessment literacy in MBT (QALMBT-Modeling Score) when it was.

Another limitation might be related to social desirability bias with the QALMBT questionnaire. Social desirability bias refers to the tendency of research subjects to provide answers that might differ from their actual attitudes, values or behaviors (Larson, 2019) in a way that is perceived as desirable by others (Kuncel & Tellegen, 2009). In this vein, it is worth noting that in the consent ISTs were informed that their answers would be anonymous, and the results would be reported by using aggregated scores to mitigate social desirability bias. Moreover, the administration of the questionnaire online was preferred to reduce ISTs' unconscious need for approval from the researcher which might have occurred in person. Another limitation might be related to the common statement used to differentiate both approaches to teaching science might have overlapped for participants. Even though I checked the survey completion times, it was not possible to know if teachers took their time to reflect on their pedagogy for each item based on a general approach to teach science or based on an MBT approach. Despite these limitations, the findings from the large sample of ISTs (43 Canadian ISTs and 373 Chilean ISTs) who participated in this study offer valuable insight into ISTs' assessment literacy in MBT based on the implementation of a new scale to identify this construct. Moreover, the analysis of class observations and interviews provided significant information to complement data from this phase of the study.

Regarding the development of assessment literacy in MBT, the data predominantly comes from a group of five Chilean ISTs. While I recognize the limitations of what we can conclude and generalize from this sample of five ISTs, I believe this study offers important information about the implementation of MBT in Chile and suggests that teachers need more assistance in MBT. As mentioned in the method chapter, to ensure trustworthiness in the qualitative analysis,

transferability, credibility, dependability and confirmability, different strategies were used. Transferability in this study was achieved by reporting the instruments, contexts, processes, and data from the cross-case analysis of five ISTs (Morrow, 2005). Therefore, a cross-cases analysis was undertaken to provide rich information about ISTs' assessment literacy, however, as with case study research, the results and conclusions might only apply to this specific context and a specific sample of teachers. Furthermore, the number of teachers interested in participating in the OPDC limited the possibility of selecting participants based on specific criteria of interest (e.g., subject, years of experience). Hence, a convenience and volunteer sample were ultimately used to identify participants in this phase of the research on assessment literacy in MBT. One of the characteristics of these type of strategies is that the sample generated may differ from the overall population because some participants might be pre-disposed to be studied (e.g., based on their own interest to enrich their pedagogy) and participate in the study (Brownell et al., 2013). Regarding volunteer sampling, for example, the data from new teachers might have impacted the quality of data since two ISTs had no more than two years of teaching experience teaching science. A random sample or a larger convenience sample could have revealed different results with the inclusion of teachers had different years of teaching experience and preparation, and potentially add to the diversity of perspectives regarding the assessment practices of ISTs when teaching with models in science. Future research would thus benefit from diverse sampling techniques.

To ensure the credibility of the study, during class observation I focused on those practices that were most relevant to answer research question 2. Moreover, methodological triangulation was pursued by using different methods of data collection that involved quantitative data from the

QALMBT questionnaire and qualitative data from interviews, class observations, observational rubric, and ISTs' artifacts. The factors evidenced from the exploratory factor analysis were then explored and enriched from the analysis of the themes that emerged from the thematic analysis. Also, codes were developed and revised constantly to analyze ISTs' pedagogy and identify the variety and frequency of ISTs' assessment practices for each theoretical dimension included in Figure 1. To ensure dependability and confirmability, each of the research steps was carefully detailed, and also a detailed description of the construction and validation of the instruments (QALMBT questionnaire and rubrics) was provided.

It is worth noting that the findings are limited by each school setting's contexts since, based on the data from the interviews, some teachers mentioned they had to align their summative assessments to the confines of their school structure and mandate. This mandate might have limited teachers' flexibility and creativity to implement new testing strategies in their pedagogy. In this sense, it cannot be assumed that teachers' limited assessment literacy in MBT was only due to their lack of knowledge and preparation in MBT but also the context might have influenced each of the theoretical dimensions included in Figure 1. More evidence and further studies are needed to support the conclusions. Furthermore, the uncontrollable aspects of the political context in Chile, such as government protests, classroom disruptions, and school closures, affected the ISTs capacity to teach and implement MBT in the science classroom. The teaching opportunities available to ISTs following these protests might have reinforced ISTs' pressure to teach the remaining curriculum content quickly through a traditional method. For example, in the interview after the OPDC, Gabriel mentioned that he regularly, and particularly after the riots, felt pressure from the director of the curriculum to continue teaching disciplinary

core ideas instead of challenging students to generate their own explanations. Additional research is needed to identify why ISTs' ideas about MBT changed after the OPDC according to interviews rather than actual observable instructional changes. First, I hypothesize that changes in ISTs' practice do not happen as quickly as changes in their thoughts about how to teach science. For example, in Vo et al.'s (2019) longitudinal study, these scholars investigated how ISTs understood and used scientific modeling to engage students in the study of the water cycle over time (3 years). They found that the enrichment of in-service science teachers' practices (e. g., capacity to identify students' modeling needs) in MBT, as a result of attending a multi-year professional development, might take years for them to be more aligned with the assessment of modeling practices. In order to reshape their pedagogy, teachers need to reflect on their practice, revisit their lesson plans, think back on their actions, and reflect on the assessment strategies implemented to assess students' models and modeling practices. In this sense, future studies need to explore why teachers' self-report was not always aligned with their pedagogy in MBT. Second, I hypothesize that ISTs enrich and reshape their assessment literacy when they teach the same curricular model every year or among different courses. In other words, with practice and more experience teaching the same lesson, ISTs can more easily identify what assessment strategies they can implement to assess a particular target model. Hence, they can judge the impact of their decisions when assessing students' models. In this sense, a more longitudinal study would be suggested to explore changes in IST's assessment practices.

Regarding the OPDC, it is not possible to conclude the full impact of the 10-hour OPDC on ISTs' assessment practices because there were different factors that influenced their participation in the course, such as their time availability, motivation, and work overload. Ideally, ISTs should

have had more time to attend the online course rather than the four weeks in which they were asked to read the course material, do assignments and revise and refine their lesson plans. Professional development experiences for science teachers are suggested to last at least 20 hours (Desimone, 2009) distributed over several weeks to help participants process new insight that allows them to reshape their instruction (Ogan-Bekiroglu, 2007). Further evidence and more research on ISTs' pedagogy would be required to make inferences about the effectiveness of the OPDC and the real impact on teachers' assessment literacy in MBT. To do so, a revised OPDC could be offered to a larger sample of teachers in the future and over a longer time interval (cf., Vo et al., 2019; Zangori et al., 2017) to allow ISTs not only to plan, use and reflect on how to use new assessment strategies to assess students' reasoning with models but also to gather evidence of their assessment practices to assess the impact of their ALMBT. Some recommendations for future OPDC drawn for this study include i) teaching ISTs how to develop assessment instruments to summatively assess students in the science classroom, ii) teaching ISTs how to develop and use scoring rubrics (see, for example, Merrit and Krajcik's (2013) study) to assess the generation, evaluation, and modification of models, iii) engaging ISTs in the co-design and critique of their own lesson plans and particularly their assessment instruments and assessment strategies to teach them how to adapt their current curriculum materials into an MBT approach (see, for example, Becker & Jacobsen, 2019; Thompson et al., 2019; Vo et al., 2019; Zangori et al., 2017), and iv) teaching ISTs how to create learning progressions and scaffolding activities to identify how students' reasoning with a model and modeling practices progress.

5.4 Significance

The significance of the findings of this study are threefold. Regarding the potential for a

theoretical contribution, this study enriches the little empirical and theoretical work examining assessment literacy. Firstly, there was not a clear definition to guide the construction of a scale to measure assessment literacy in MBT in the field of science education. Based on a definition of assessment literacy from a general approach of teaching, I adapted this conceptualization to include the foundations of MBT. MBT is expanding as the approach to teaching science (Buckley et al., 2004; Chiu & Lin, 2019; Clement, 2000; Lehrer & Schauble, 2010; Windschitl et al., 2008; Zangori et al., 2017) and this study represented a novel focus on ISTs' assessment literacy within school settings and contexts. The results of this study based on the analysis of the QALMBT and the themes identified from the qualitative data based on the analysis of interviews, classroom observations and ISTs' artifacts revealed that ALMBT includes three major components which teachers need to put into practice when assessing students' models and modeling. These components or dimensions include that ISTs must i) have a sophisticated knowledge of the implementation of strategies to promote the elicitation and assessment of student's models, ii) promote self and peer assessment of generated models, and iii) design and implement assessment criteria and assessment instruments to assess students' reasoning about generated models. Future research should be carried out to confirm these dimensions and establish how each of these components of assessment literacy interacts and shapes teachers' decisions when developing and implementing specific assessment practices and assessment instruments. In chapter 2, assessment literacy in MBT (ALMBT) was defined as a multidimensional construct that is comprised of a set of knowledge and skills about the assessment of models and modeling which is activated in the science classroom when reflecting on practice while students generate, evaluate and modify their initial models. Although evidence of a sophisticated variety of assessment strategies for each of the three suggested dimensions

were not observed among ISTs, the dimensions suggested by the exploratory factor analysis and then explored by the analysis of qualitative data lead to an operationalization of ALMBT and offer guidelines for the elements that ISTs need more support to assess students' models. Future studies must focus on confirming or adapting the definition of ALMBT and the three suggested dimensions and also explore what variables such as years of experience, have major influences on helping ISTs to transition into higher levels of proficiency in assessment in MBT.

From a *methodological perspective*, the QALMBT questionnaire is a new instrument to identify ISTs' assessment literacy in MBT. I collected validity and reliability evidence to justify the psychometric properties of the QALMBT questionnaire. This evidence was provided in order to support the conclusions that were used to answer the first research question about whether ISTs' knowledge of models and modeling was related to their assessment literacy in MBT. The high value of the omega coefficient provided evidence of the internal consistency of the questionnaire. The results suggest that ISTs' frequency of assessment strategies in MBT is related to how much they know about the nature of models in science and other predictors such as years of teaching experience, the number of topics studied in science, and the number of courses taken in assessment. Finally, the limited body of existing literature in assessment literacy in MBT does not include a measure of ISTs' modeling assessment. The results of the exploratory factor analysis provided evidence of a three-dimensional ALMBT scale. Moreover, the observation rubric (R-ASMM) and R-LPAL are new instruments that can be used to characterize ISTs' modeling-based assessment practice in the science classroom. These rubrics can also be used as a framework of teacher education programs or science methodological courses to enrich ISTs'

assessment literacy in MBT and offer a repertoire of strategies that future and current science teachers can implement to assess students' models and modeling practices.

Finally, from a *practical* point of view, the results of this study provide a measure of assessment literacy in MBT that has been utilized successfully in this study in two different countries. The construction and administration of the QALMBT questionnaire is the first step in continuing to conduct studies that might help us to identify teachers' assessment strategies when engaging students in modeling practices. Moreover, the results of this study offer a better understanding of the underpinnings of assessment literacy of science teachers, and the results might be useful for stakeholders in science education and especially for curriculum developers, many of whom must consider teachers' preparedness for MBT. The findings of this research further showed that teachers selected and organized their lesson plans based on the national curriculum, but they lacked preparation and training when developing and implementing assessment strategies that require the construction and refinement of models even though they are explicitly indicated as learning goals in the prescribed curriculum. In this sense, it is important for curriculum developers and teacher educators to explicitly guide ISTs regarding modeling practices they need to teach to their students through the GEM cycle. This outcome can be achieved through greater guidance and helping teachers to develop and implement formative and summative assessments to assess students' progress when thinking with models (e.g., by suggesting indicators of assessment that explicitly cover each phase involved in MBT). For example, the construction of a portfolio by students that includes the collection of initial and intermediate models generated by students might help ISTs identify how students' understanding of a model has been enriched within a unit and might also support students' reflection on their modeling practices. Another

example can be related to encouraging students to evaluate different scenarios to challenge them to explore the utility, scope, and limitation of a model by using assessment criteria related to good models. Also, professional development programs should place firm attention on enriching ISTs' epistemological knowledge of models and modeling in science and also on providing teachers with examples of pedagogical tools to exemplify how teachers can assess the construction, evaluation, and modification of models in the science classroom. Many ISTs acknowledged the importance of scientific models in science, but based on their responses in the interviews, they appeared to lack experience and opportunities to create, implement, and refine assessment tools to evaluate students' modeling practices. A probable explanation is that teachers' proficiency levels in ALMBT might be connected to their years of experience, and disciplinary knowledge and PCK, as suggested by Samantha's case and based on the analysis of the regression models of the QALMBT. It is worth mentioning that ISTs' acquisition of assessment knowledge is not a process that occurs linearly. Instead, it likely requires a long-term commitment with a close process of coaching and mentoring from the teacher educator (Mak, 2019). In this sense, given the context-specific nature of an assessment in MBT, future OPDC must immerse ISTs in more specific activities on how to develop assessment instruments to enhance ISTs' knowledge of assessment and their repertoire to assess students' models and modeling practices. Moreover, future research must include a larger sample of ISTs and control for the variable related to years of teaching experience to explore how this variable might influence the complexity of ISTs' assessment strategies in the classroom. Moreover, the comparison among different disciplines, such as chemistry, physics and biology, might be another factor to consider in future studies in order to identify if ISTs assess their students differently based on the discipline and content they teach.

5.5 Recommendations for Practice and Further Studies

The findings of this study also open the way to identify ISTs' strengths and suggest what aspects of the dimensions related to their assessment literacy in MBT need to be enriched. Four categories of recommendations for researchers and practitioners emerged from this study. Each of the categories are indicated and detailed below.

5.5.1 Examine and Revise the Psychometric Properties of the QALMBT

Future investigations might examine the factor structure in order to conduct a confirmatory factor analysis. The stability of the factors identified from the EFA needs to be studied to support the three-factor structure of the QALMBT-Modeling and evaluate the model fit. Cross-cultural analysis with samples from different contexts should be conducted in the future since many constructs in psychology and education vary given cultural differences that shape participants beliefs, practices, social roles and norms, and organizational structures (Ilesanmi, 2009). This analysis might be important since during the interviews and private conversations some ISTs expressed the fact that their opportunities to innovate in the implementation of summative assessment strategies were limited and determined by schools. In this sense, for example, for the factor related to the design and implementation of assessment criteria and assessment instruments to assess students' reasoning about generated models, I might hypothesize that some ISTs might prioritize the development of assessment instruments that are aligned to standardized exams suggested by each school setting instead of assessing summatively students' modeling practices. Furthermore, regarding the linear regression conducted to answer the first research question, further studies might be conducted to explore group-level clustering by using

multilevel modeling. I acknowledge that the use of ordinary least squares (OLS) regression might increase the risk of Type I error (concluding that there are significant effects when they might have occurred by chance). Nevertheless, it is worth mentioning that the majority of the predictors included in the linear regression were significant at p-values smaller than .01 which minimizes significantly the risk of obtaining results that contain a type I error. Multilevel modeling might be useful to identify if there are differences among ISTs' assessment literacy self-report based on their demographic information (e. g., region). In other words, "Multilevel models appropriately partition within-group and between group effects so that a high level of clustering within groups is statistically accounted for" (Clarke, 2008, p. 752).

5.5.2 Refinement of the OPDC and New Opportunities for Professional Development

While not the focus of analysis in this study, the OPDC that I developed provided ISTs with general insight about how to implement MBT in the science classroom and about how to assess students when thinking with models. This OPDC attempted to support ISTs in MBT by providing them with the main foundations of this approach and guiding them to reflect on how they use assessment to engage students in thinking with models. Regardless, the OPDC in MBT was not intended to be a course in assessment. Hence, teachers might have struggled to enrich their assessment practices in MBT even after reading the modules. It is recommended that future studies place more attention on enriching each of the theoretical dimensions related to ALMBT and investigating them. To improve the impact of the OPDC, another strategy might be restructuring the course and asking ISTs to attend an in-person series of workshops in MBT in which ISTs are required to cover each of the modules over a longer span of time. Through such professional development activities, ISTs might interact with their colleagues and create teacher

learning communities in which they can review, criticize, reflect, and enrich their pedagogy, lesson plans, and assessment instruments together and synchronously with the instructor (see, for example, Brady et al., 2011; Bridle & Yezierski, 2011; Guy-Gaytán et al., 2019; Merritt & Krajcik, 2013). Another suggestion might include the creation of an online professional development through virtual learning communities in which ISTs co-construct their knowledge for each of the dimensions of assessment literacy by sharing, providing feedback and commenting on each others' lesson plans, assessment instruments, and assessment strategies developed to assess students' models and modeling practice.

5.5.3 Rethinking Teacher Preparation

Even though my study did not explore pre-service science teachers' assessment literacy in MBT, the case of Lisa and Gabriel are informative regarding science teacher preparation. Both ISTs had less than 2 years of teaching experience and, in the case of Lisa, she had just finished her degree. The limited repertoire in assessment in MBT of these two ISTs might suggest that science teacher preparation is limited in ALMBT. In this sense, science teacher programs should be aware of these findings and support pre-service science teachers as well on MBT in the science classroom and how to engage in summative and formative assessments. Limitations in ISTs' preparation may not be unique to the sample of ISTs who participated in this study. Although ISTs did not have a rich repertoire in MBT, it may be that they were beginning to think about the relevance of modeling practices in their pedagogy that would have allowed them to enrich their instruction in the future. Future studies might expand the discussion on the possible reasons why ISTs struggle to incorporate assessment practices in MBT in their pedagogy. Based on the results of this study, I recommend that science methods courses cover content not only

related to the disciplinary knowledge of models and modeling in science education but specifically teach pre-service science teachers i) about assessment strategies that can promote students' engagement in modeling activities through the elicitation of their models; ii) about how to assess each phase of the GEM cycle and particularly the evaluation and modification phase by engaging students in peer and self-assessment; iii) about how to create and adapt assessment scoring tools into an MBT approach; iv) about how to use feedback to help students revise their models; and v) about how to interpret the results from assessment to reshape their pedagogy in order to facilitate students' progression in their reasoning with models and modeling practices. It is also worth mentioning that pre-service teachers need to be exposed early to MBT experiences during their teacher education programs to help them understand the foundations of model-based inquiry in the science classroom and acquire experience in the assessment of models in science education.

5.5.4 Supporting ISTs with a Sophisticated Repertoire for Assessing Science Inquiry

The majority of participants in this study lacked sophisticated repertoires to engage students in class during the evaluation and modification phases of the GEM cycle. These phases were mostly absent in many of the lessons even after the OPDC. In this sense, teachers might need further guidance strategies to help them embrace the epistemology of modeling and how to assess students' modeling practices beyond what the OPDC was able to provide. I observed teachers formulating driving questions not only to assess and judge students' understanding but also to challenge students to clarify and support their claims. The ISTs followed a linear pattern that focused on asking a question, providing feedback, giving the correct answer, and asking a new factual question. I did not observe assessment strategies that might have challenged students

to be involved in assessment while thinking with models, such as assessing or judging their peers' ideas, testing hypotheses that were generated, or comparing assumptions or components between two or more models. Also, data from interviews revealed that ISTs possessed a general idea about how to design summative assessment tools but did not explicitly provide evidence about how they might assess students' alternative ideas or what elements they might include when designing specific scoring tools to evaluate students' generated and revised models. The findings of this study also showed that the learning of a target model was, on many occasions, rushed. ISTs did not give enough opportunities to their students to think and use models in the classroom, and they focused on the retention and understanding of disciplinary ideas rather than helping students progress in their learning of models. In this sense, teachers seem to need to be supported in i) how to identify and anticipate intermediate and alternative models that students might express in the classroom; ii) how to formulate driving questions that might guide students in the process of enrichment and modification of their models and iii) supporting ISTs' discourse in modeling to engage students in thinking with models before teaching them the expected curricular or target model or the underlying mechanisms or process related to some phenomenon.

Pedagogical assistance for ISTs when working with large class size must also be explored. The impact of class-size on the ability of ISTs to implement MBT in the classroom might be a factor that limits ISTs' capacity to assess students' generated models. In the case of Samantha, she worked with a small group of students and had more opportunities to guide and monitor each student when thinking with models. As a recommendation, when ISTs do not have enough time in their class or teach large classes, I suggest ISTs must focus on i) identifying the most important aspects or mechanism of the target model rather than trying to cover large amount of

content information that is irrelevant for students; ii) clearly and explicitly informing students the learning objectives for each lesson by emphasizing the model that is students are expected to generate and the modeling practices that each specific activity is attempting to promote;, iii) starting gradually the transition from a lecture-based approach of teaching to a more active learning by giving students more responsibilities in the classroom and involving them in the generation, evaluation and modification of their models; iv) using small groups to challenge students to create models and then compare and evaluate them in the class until reaching a consensus model; v) asking randomly students to elicit their ideas about a model in order to identify students' alternative ideas and use them to promote conceptual change in the class; vi) using technology such as web-based software to challenge students to generate hypotheses and manipulate variables of a model; and vii) designing scaffolded activities that help students be aware of how their modeling practices have been enriched, for example, within a unit.

Finally, it is worth noting that this study did not focus on students' performance beyond the data that was obtained during teacher-student interaction in the class observations. Further studies might focus on exploring the impact of ISTs' assessment literacy in MBT on students' learning with models, and analyzing which component of the ISTs' assessment literacy (e. g., knowledge of peer and self-assessment, knowledge of grading) might have a bigger impact on students' achievement and modeling-related skills.

References

Abdalla, M. M., Oliveira, L. G. L., Azevedo, C. E. F., & Gonzalez, R. K. (2018). Quality in qualitative organizational research: Types of triangulation as a methodological alternative. *Administração: Ensino e Pesquisa, 19*(1), 66-98. https://doi.org/10.13058/raep.2018.v19n1.578

Abdi, H. (2003). Partial regression coefficients. In M. Lewis-Beck, A. Bryman, T. Futing (Eds.), *Encyclopedia for research methods for the social sciences* (pp. 978-982). Sage.

Abell, S. K., & Siegel, M. A. (2011). Assessment literacy: What science teachers need to know and be able to do. In D. Corrigan, J. Dillon & R. Gunstone (Eds.), *The professional knowledge base of science teaching* (pp. 205-221). Springer.

Afari, E. (2015). Relationships of students' attitudes toward science and academic achievement. In M. S. Khine (Ed.), *Attitude measurements in science education: Classic and contemporary approaches* (pp. 245-262). IAP.

Allen, M. (2017). *The SAGE encyclopedia of communication research methods*. SAGE.

Allison, P. D. (1999). *Multiple regression: A primer*. Pine Forge Press.

Alonzo, A. C. (2018). An argument for formative assessment with science learning progressions. *Applied Measurement in Education, 31*(2), 104-112. https://doi.org/10.1080/08957347.2017.1408630

Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. *Science Education, 93*(3), 389-421. https://doi.org/10.1002/sce.203

American Federation of Teachers, National Council on Measurement in Education, & National Education Association (AFT, NCME, & NEA). (1990). Standards for teacher competence

in educational assessment of students. *Educational Measurement: Issues and Practice, 9*(4), 30-32. https://doi.org/10.1111/j.1745-3992.1990.tb00391.x

Anney, V. N. (2014). Ensuring the quality of the findings of qualitative research: Looking at trustworthiness criteria. *Emerging Trends in Educational Research and Policy Studies, 5*(2), 272-281.

Areepattamannil, S., Cairns, D., & Dickson, M. (2020). Teacher-directed versus inquiry-based science instruction: Investigating links to adolescent students' science dispositions Across 66 Countries. *Journal of Science Teacher Education*, *31*(6), 675-704. https://doi.org/10.1080/1046560X.2020.1753309

Asghar, J., & Ahmad, A. (2014). Teacher Development: An overview of the concept and approaches. *Journal of Educational and Social Research, 4*(6), 147-160. https://doi.org/10.5901/jesr.2014.v4n6p147

Atkinson, T. M., Sit, L., Mendoza, T. R., Fruscione, M., Lavene, D., Shaw, M., Li, Y., Hay, J., Cheeland, C. S., Scher, H. I., Breitbart, W. S., & Basch, E. M. (2010). Confirmatory factor analysis to evaluate construct validity of the Brief Pain Inventory (BPI). *Journal of Clinical Oncology, 28*(15), 558-565. http://doi.org/10.1016/j.jpainsymman.2010.05.008

Avalos, B. (2011). Teacher professional development in teaching and teacher education over ten years. *Teaching and Teacher Education, 27*(1), 10-20. https://doi.org/10.1016/j.tate.2010.08.007

Bandele S. O., & Oluwatayo, J. A. (2013). Assessing Assessment Literacy of Science Teachers in Public Secondary Schools in Ekiti State. *Journal of Education and Practice, 4*(28), 56-63.

Battaglia, M. (2011). Convenience sampling. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (p. 149). SAGE.

Bayar, A. (2014). The Components of effective professional development activities in terms of teachers' perspective. *International Online Journal of Educational Sciences, 6*(2), 319-327. http://dx.doi.org/10.15345/iojes.2014.02.006

Beaton, D.E., Bombardier, C., Guillemin, F., & Ferraz, M.B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine, 25*(24), 3186–3191. http://doi.org/10.1097/00007632-200012150-00014

Beck, J. P., Muniz, M. N., Crickmore, C., & Sizemore, L. (2020). Physical chemistry students' navigation and use of models to predict and explain molecular vibration and rotation. *Chemistry Education Research and Practice, 21*(2), 597-607. https://doi.org/10.1039/C9RP00285E

Behr, D. (2017). Assessing the use of back translation: The shortcomings of back translation as a quality testing method. *International Journal of Social Research Methodology, 20*(6), 573-584.

Bell, B. (2007). Classroom assessment of science learning. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 965-1006). Lawrence Erlbaum.

Bellei, C. (2009). The public–private School controversy in Chile. In R. Chakrabarti & P. E. Peterson (Eds.), *School Choice International: exploring public–private partnerships* (pp. 165-192). MIT Press.

Ben-Nun, P. (2008). Respondent fatigue. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 743-744). SAGE.

Bennett, R. E. (2011). Formative assessment: a critical review. *Assessment in Education: Principles, Policy & Practice, 18*, 5–25. http://doi.org/10.1080/0969594X.2010.513678

Bennett, S. C. (2017). *Adapting a framework for assessing students' approaches to modeling* (Publication No. 10618843) [Doctoral dissertation, Michigan State University]. ProQuest Dissertations Publishing.

Besnoy, K. D., Dantzler, J., Besnoy, L. R., & Byrne, C. (2016). Using exploratory and confirmatory factor analysis to measure construct validity of the Traits, Aptitudes, and Behaviors Scale (TABS). *Journal for the Education of the Gifted, 39*(1), 3-22. http://doi.org/10.1177/0162353215624160

Black, P., & Wiliam, D. (1998). *Inside the black box*. King's College.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5. https://doi.org/10.1007/s11092-008-9068-5

Black, P., Harrison, C., & Lee, C. (2003). *Assessment for learning: Putting it into practice*. Open University Press.

Blair, E. (2015). A reflexive exploration of two qualitative data coding techniques. *Journal of Methods and Measurement in the Social Sciences, 6(*1), 14-29. https://doi.org/10.2458/v6i1.18772

Boekaerts, M., & Corno, L. (2005). Self-regulation in the classroom: A perspective on assessment and intervention. *Applied Psychology, 54*(2), 199-231. http://10.1111/j.1464-0597.2005.00205.x

Borg, S. (1998). Data-based teacher development. *English Language Teaching Journal, 52*(4), 273-281. https://doi.org/10.1093/elt/52.4.273

Borko, H., & Putnam, R. (1996). Learning to teach. In D. Berliner, & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 673-708). Macmillan International Higher Education.

Boud, D. (2000). Sustainable assessment: rethinking assessment for the learning society. *Studies in Continuing Education, 22*(2), 151-167. https://doi.org/10.1080/713695728

Brace, N., Snelgar, R., & Kemp, R. (2012). *SPSS for Psychologists*. Macmillan International Higher Education.

Bradford, S., & Cullen, F. (2012). *Research and research methods for youth practitioners*. Routledge.

Brady, C., Holbert, N., Soylu, F., Novak, M., & Wilensky, U. (2015). Sandboxes for model-based inquiry. *Journal of Science Education and Technology, 24*(2-3), 265-286. http://doi.org/10.1007/s10956-014-9506-8

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*, 77–101. http://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2013). *Successful qualitative research: A practical guide for beginners*. Sage.

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society*, 1287-1294. https://doi.org/10.2307/1911963

Bridle, C. A., & Yezierski, E. J. (2011). Evidence for the effectiveness of inquiry-based, particulate-level instruction on conceptions of the particulate nature of matter. *Journal of Chemical Education, 89*(2), 192-198. https://doi.org/10.1021/ed100735u

British Columbia Ministry of Education (2018a). *Core competencies*. Retrieved from the Ministry of Education of British Columbia website: https://curriculum.gov.bc.ca/competencies

British Columbia Ministry of Education (2018b). *Critical thinking*. *Competency profiles*. Retrieved from the Ministry of Education of British Columbia website: https://curriculum.gov.bc.ca/sites/curriculum.gov.bc.ca/files/pdf/CriticalThinkingCompetencyProfiles.pdf

British Columbia Ministry of Education (2018c). *Area of learning: Science curriculum – anatomy and physiology Grade 12*. Retrieved from the Ministry of Education of British Columbia website: https://curriculum.gov.bc.ca/curriculum/Science

Brown, G. T. L. (2008). *Conceptions of assessment: Understanding what assessment means to teachers and students*. Nova Science.

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice, 30*(1), 3-12. https://doi.org/10.1111/j.1745-3992.2010.00195.x

Brownell, S. E., Kloser, M. J., Fukami, T., & Shavelson, R. J. (2013). Context matters: volunteer bias, small sample size, and the value of comparison groups in the assessment of research-based undergraduate introductory biology lab courses. *Journal of Microbiology & Biology Education, 14*(2), 176. http://doi.org/10.1128/jmbe.v14i2.609

Buckley, B. C., Gobert, J. D., Kindfield, A. C. H., Horwitz, P., Tinker, R. F., Gerlits, B., Wilensky, U., Dede, C., & Willett, J. (2004). Model-based teaching and learning with BioLogicaTM: What do they learn? How do they learn? How do we know? *Journal of*

Science Education and Technology, 13(1), 23–41. https://doi.org/10.1023/B:JOST.0000019636.06814.e3

Cabalin, C. (2012). Neoliberal education and student movements in Chile: Inequalities and malaise. *Policy Futures in Education, 10*(2), 219-228. https://doi.org/10.2304/pfie.2012.10.2.219

Campbell, S. (2016). Perspectives: Method and methodology in nursing research. *Journal of Research in Nursing, 21*(8), 656-659. https://doi.org/10.1177/1744987116679583

Campbell, T., Zhang, D., & Neilson, D. (2011). Model based inquiry in the high school physics classroom: An exploratory study of implementation and outcomes. *Journal of Science Education and Technology, 20*(3), 258-269. http://doi.org/10.1007/s10956-010-9251-6

Carless, D. (2006). Differing perceptions in the feedback process. *Studies in Higher Education 31*(2), 219–233. https://doi.org/10.1080/03075070600572132

Carlson, J., & Daehler, K. (2019, 2019). The refined consensus model of pedagogical content knowledge in science education. In A. Hume, R. Cooper, & A. Borowski (Eds.), *Repositioning pedagogical content knowledge in teacher' knowledge for teaching science* (pp. 93–116). Springer.

Carr, M., McGee, C., Jones, A., McKinley, E., Bell, B., Barr, H., & Simpson, T. (2000). *Strategic research: Initiative literature review: The effects of curricula and assessment on pedagogical approaches and on educational outcomes*. Ministry of Education of New Zealand.

Casson, R. J., & Farmer, L. D. (2014). Understanding and checking the assumptions of linear regression: A primer for medical researchers. *Clinical & Experimental Ophthalmology, 42*(6), 590-596. http://doi.org/10.1111/ceo.1235

Castillo-Montoya, M. (2016). Preparing for interview research: The interview protocol refinement framework. *The Qualitative Report, 21*(5), 811-831. http://doi.org/10.46743/2160-3715/2016.2337

Castleberry, A., & Nolen, A. (2018). Thematic analysis of qualitative research data: Is it as easy as it sounds? *Currents in Pharmacy Teaching and Learning, 10*(6), 807-815. http://10.1016/j.cptl.2018.03.019

Cattell, R.B. (1973). *Factor analysis*. Greenwood Press.

Chan, K. K. H., & Yung, B. H. W. (2018). Developing pedagogical content knowledge for teaching a new topic: More than teaching experience and subject matter knowledge. *Research in Science Education, 48*(2), 233-265. https://doi.org/10.1007/s11165-016-9567-1

Chang, S. N. (2007). Externalising students' mental models through concept maps. *Journal of Biological Education, 41*(3), 107-112. https://doi.org/10.1080/00219266.2007.9656078

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. SAGE.

Cheng, M. F., & Brown, D. E. (2015). The role of scientific modeling criteria in advancing students' explanatory ideas of magnetism. *Journal of Research in Science Teaching, 52*(8), 1053-1081. https://doi.org/10.1002/tea.21234

Cheng, M. F., & Lin, J. L. (2015). Investigating the relationship between students' views of scientific models and their development of models. *International Journal of Science Education, 37*(15), 2453-2475. https://doi.org/10.1080/09500693.2015.1082671

Chichekian, T., Shore, B. M., & Tabatabai, D. (2016). First-year teachers' uphill struggle to implement inquiry instruction: Exploring the interplay among self-efficacy,

conceptualizations, and classroom observations of inquiry enactment. *SAGE Open, 6*(2), 1-19. https://doi.org/10.1177/2158244016649011

Chin, C., & Brown, D. E. (2000). Learning in science: A comparison of deep and surface approaches. *Journal of Research in Science Teaching, 37*(2), 109-138. https://doi.org/10.1002/(SICI)1098-2736(200002)37:2%3C109::AID-TEA3%3E3.0.CO;2-7

Chiu, M.-H., & Lin, J.-W. (2019). Modeling competence in science education. *Disciplinary and Interdisciplinary Science Education Research, 1*(12), 1-11. https://doi.org/10.1186/s43031-019-0012-y

Chun Tie, Y., Birks, M., & Francis, K. (2019). Grounded theory research: A design framework for novice researchers. *SAGE Open Medicine, 7*, 1-8. https://doi.org/10.1177/2050312118822927

Clarke, V., Braun V., & Hayfield, N. (2015). Thematic Analysis. In J. A. Smith, (Ed.), *Qualitative psychology: A practical guide to research methods* (pp. 222-248). Sage.

Clarke, P. (2008). When can group level clustering be ignored? Multilevel models versus single-level models with sparse data. *Journal of Epidemiology & Community Health, 62*(8), 752-758.

Clement, J. (1989). Learning via model construction and criticism. In J. Glover, C. Reynolds, & R. Royce (Eds.), *Handbook of creativity* (pp. 341–381). Springer.

Clement, J. (1993). Using bridging analogies and anchoring intuitions to deal with students' preconceptions in physics. *Journal of Research in Science Teaching, 30*(10), 1241–1257. https://doi.org/10.1002/tea.3660301007

Clement, J. (2000). Model based learning as a key research area for science education. *International Journal of Science Education, 22*(9), 1041-1053. https://doi.org/10.1080/095006900416901

Clement, J. J., & Rea-Ramirez, M. A. (2008). *Model-based learning and instruction in science*. Springer.

Coll, R. K., & Lajium, D. (2011). Modeling and the future of science learning. In M. S. Khine & I. M. Saleh (Eds.), *Models and modeling* (pp. 3-21). Springer.

Coll, R. K., France, B., & Taylor, I. (2005). The role of models/and analogies in science education: implications from research. *International Journal of Science Education, 27*(2), 183-198. https://doi.org/10.1080/0950069042000276712

Contreras, S. (2010). *Las creencias y actuaciones curriculares de los profesores de ciencias de secundaria de Chile* [Beliefs and curricular actions of secondary teachers from Chile] (Doctoral dissertation). Retrieved from https://educar.ec/jornada/saul%20alejandro%20contreras%20palma.pdf

Contreras, S. A. (2016). Pensamiento Pedagógico en la Enseñanza de las Ciencias: análisis de las creencias curriculares y sus implicancias para la formación de profesores de enseñanza media [Pedagogical thinking in science education. Analysis of curricular beliefs and their implications for high school teacher training]. *Formación Universitaria*, *9*(1), 15-24. http://dx.doi.org/10.4067/S0718-50062016000100003

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics, 19*(1), 15-18. https://doi.org/10.2307/1268249

Corbin, J. M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology, 13*(1), 3-21. https://doi.org/10.1007/BF00988593

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation, 10*(1), 1-9. https://doi.org/10.7275/jyj1-4868

Covitt, B. A., Gunckel, K. L., Caplan, B., & Syswerda, S. (2018). Teachers' use of learning progression-based formative assessment in water instruction. *Applied Measurement in Education, 31*(2), 128-142. https://doi.org/10.1080/08957347.2017.1408627

Crawford, B., & Cullin, M. (2004). Supporting prospective teachers' conceptions of modeling in science. *International Journal of Science Education, 26*, 1379–1401. https://doi.org/10.1080/09500690410001673775

Crawford, B., & Cullin, M. (2005). Dynamic assessments of pre-service teachers' knowledge of models and modelling. In K. Boersma, H. Eijkelhof, M. Goedhart, & O. Jong (Eds.), *Research and the quality of science education* (pp. 309–323). Springer.

Creswell, J. W. (2010). Mapping the developing landscape of mixed methods research. In A. Tashakkori & C. Teddie (Eds.), *Handbook of mixed methods in social & behavioral research* (2nd ed.) (pp. 45-68). Sage.

Creswell. J. W. & Plano Clark, V. L. (2011). Designing and conducting mixed methods research (2nd ed.). Sage.

Creswell, J. W., Plano Clark, V. L., Gutmann, M. L., & Hanson, W. E. (2003). Advanced mixed methods research designs. In A. Tashakkori, & C. Teddlie (Eds.), *Handbook of mixed Methods in social and behavioral research* (pp. 209-240). SAGE.

Crompton, H., Burke, D., Gregory, K. H., & Gräbe, C. (2016). The use of mobile learning in science: A systematic review. *Journal of Science Education and Technology, 25*(2), 149-160. https://doi.org/10.1007/s10956-015-9597-x

Cutcliffe, J. R., & McKenna, H. P. (2004). Expert qualitative researchers and the use of audit trails. *Journal of Advanced Nursing, 45*(2), 126-133. https://doi.org/10.1046/j.1365-2648.2003.02874.x

Danczak, S. M., Thompson, C. D., & Overton, T. L. (2020). Development and validation of an instrument to measure undergraduate chemistry students' critical thinking skills. *Chemistry Education Research and Practice, 21*(1), 62-78. https://doi.org/10.1039/C8RP00130H

Danusso, L., Testa, I., & Vicentini, M. (2010). Improving prospective teachers' knowledge about scientific models and modelling: Design and evaluation of a teacher education intervention. *International Journal of Science Education, 32*(7), 871-905. https://doi.org/10.1080/09500690902833221

Dawson, J. (2012). Thick Description. In A. J. Mills, G. Durepos & E. Wiebe (Eds.), *Encyclopedia of case study research* (pp. 943-944). SAGE.

de Vaus, D. (2001). *Research design in social research*. Sage.

De Winter, J. C., & Dodou, D. (2012). Factor recovery by principal axis factoring and maximum likelihood factor analysis as a function of factor pattern and sample size. *Journal of Applied Statistics, 39*(4), 695-710. https://doi.org/10.1080/02664763.2011.610445

DeBarger, A. H., Penuel, W. R., Harris, C. J., & Kennedy, C. A. (2016). Building an assessment argument to design and use next generation science assessments in efficacy studies of

curriculum interventions. *American Journal of Evaluation, 37*(2), 174-192. https://doi.org/10.1177/1098214015581707

Delandshere, G., & Jones, J. H. (1999). Elementary teachers' beliefs about assessment in mathematics: A case of assessment paralysis. *Journal of Curriculum and Supervision, 14*(3), 216–240.

DeLuca, C., Chavez, T., & Cao, C. (2013). Establishing a foundation for valid teacher judgement on student learning: The role of pre-service assessment education. *Assessment in Education: Principles, Policy & Practice, 20*(1), 107-126. https://doi.org/10.1080/0969594X.2012.668870

Derry, S. J., Pea, R. D., Barron, B., Engle, R. A., Erickson, F., Goldman, R., Hall, R., Koschmann, T., Lemke, J., Sherin, M., & Sherin, B. L. (2010). Conducting video research in the learning sciences: Guidance on selection, analysis, technology, and ethics. *The Journal of the Learning Sciences, 19*(1), 3-53. https://doi.org/10.1080/10508400903452884

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher, 38*, 181–199. https://doi.org/10.3102/0013189X08331140

Develaki, M. (2016). Key-aspects of scientific modeling exemplified by school science models: Some units for teaching contextualized scientific methodology. *Interchange, 47(*3), 297-327. http://doi.org/10.1007/s10780-016-9277-7

Devetak, I., Vogrinc, J., & Glažar, S. A. (2009). Assessing 16-year-old students' understanding of aqueous solution at submicroscopic level. *Research in Science Education, 39*(2), 157-179. https://doi.org/10.1007/s11165-007-9077-2

Dodgson, J. E. (2019). Reflexivity in qualitative research. *Journal of Human Lactation, 35*(2), 220-222. https://doi.org/10.1177/0890334419830990

Dunn, T. J., Baguley, T., & Brunsden, V. (2013). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399–412. http://doi.org/10.1111/bjop.12046

Dwivedi, Y. K. (Ed.). (2007). Development of survey instrument: Confirmatory survey. In Y. K. (Ed.), *Consumer adoption and usage of broadband* (117-134). IGI Global.

Edwards, F. (2016). A rubric to track the development of secondary pre-service and novice teachers' summative assessment literacy. *Assessment in Education: Principles, Policy & Practice, 24*(2), 205-227. https://doi.org/10.1080/0969594X.2016.1245651

Egan, O., & Archer, P. (1985). The accuracy of teachers' ratings of ability: A regression model. *American Educational Research Journal, 22*(1), 25–34. https://doi.org/10.3102/00028312022001025

Elo, S., Kääriäinen, M., Kanste, O., Pölkki, T., Utriainen, K., & Kyngäs, H. (2014). Qualitative content analysis: A focus on trustworthiness. *SAGE Open, 4*(1), 1-10. http://doi.org/10.1177/2158244014522633

Enders, C. K. (2006). Analyzing structural equation models with missing data. In G. R. Handcock & R.O. Mueller (Eds.), *Structural equation modeling: A second course* (p.p. 313-342). Information Age Publishing.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods, 4*(3), 272. https://doi.org/10.1037/1082-989X.4.3.272

Faraway, J. J. (2014). *Linear models with R*. CRC press.

Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods, 5*(1), 80-92. https://doi.org/10.1177/160940690600500107

Fernández, E. C. (2009). Rethinking the role of elite private schools in a neoliberal era: an example from Chile. *Policy Futures in Education, 7*(1), 31-43. http://dx.doi.org/10.2304/pfie.2009.7.1.31

Field, A. (2009). *Discovering Statistics using SPSS*. Sage.

Flaherty, J. A., Gaviria, F. M., Pathak, D., Mitchell, T., Wintrob, R., Richman, J. A., & Birz, S. (1988). Developing instruments for cross-cultural psychiatric research. *Journal of Nervous and Mental Disease, 176*(5), 257-263.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (2nd ed.). Wiley series in probability and statistics.

Flick, U. (2014). *An introduction to qualitative research*. Sage.

Fredericks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*, 59–109. https://doi.org/10.3102/00346543074001059

Friedrichsen, P. J., Abell, S. K., Pareja, E. M., Brown, P. L., Lankford, D. M., & Volkmann, M. J. (2009). Does teaching experience matter? Examining biology teachers' prior knowledge for teaching in an alternative certification program. *Journal of Research in Science Teaching, 46*(4), 357-383. https://doi.org/10.1002/tea.20283

Furtak, E. M., & Heredia, S. C. (2014). Exploring the influence of learning progressions in two teacher communities. *Journal of Research in Science Teaching, 51*(8), 982–1020. http://dx.doi.org/10.1002/tea.21156

Furtak, E. M., Thompson, J., Braaten, M., & Windschitl, M. (2012). Learning progressions to support ambitious teaching practices. In A. C. Alonzo & A. Wotwals (Eds.), *Learning progressions in science* (pp. 405-433). Sense Publishers.

Gan, Z., Nang, H., & Mu, K. (2018). Trainee teachers' experiences of classroom feedback practices and their motivation to learn. *Journal of Education for Teaching, 44*(4), 505-510. https://doi.org/10.1080/02607476.2018.1450956

Garcés, M. (2019). October 2019: Social uprising in neoliberal Chile. *Journal of Latin American Cultural Studies, 28*(3), 483-491. https://doi.org/10.1080/13569325.2019.1696289

Gasson, S. (2004). Rigor in grounded theory research: An interpretive perspective on generating theory from qualitative field studies. In Whitman, M. & Woszczynski, A. (Eds.), *The Handbook of information systems research* (pp. 79–102). IGI.

Giere, R., Bickle, J., & Mauldin, R. (2006). *Understanding scientific reasoning*. Thomson Learning.

Giere, R. N. (2004). How models are used to represent reality. *Philosophy of science*, *71*(5), 742-752. https://doi.org/10.1086/425063

Gilbert, J. K. (2004). Models and modelling: Routes to more authentic science education. *International Journal of Science and Mathematics Education, 2*(2), 115-130. https://doi.org/10.1007/s10763-004-3186-4

Gilbert, J. K., & Justi, R. (2016). *Modelling-based Teaching in Science Education*. Springer.

Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment*. The Falmer Press.

Glaser, B. G., & Strauss, A. L. (2006). *The discovery of grounded theory: Strategies for qualitative research*. Aldine Transaction.

Gobert, J. D., & Buckley, B. C. (2000). Introduction to model-based teaching and learning in science education. *International Journal of Science Education, 22*(9), 891-894. https://doi.org/10.1080/095006900416839

Gobert, J. D., O'Dwyer, L., Horwitz, P., Buckley, B. C., Levy, S. T., & Wilensky, U. (2011). Examining the relationship between students' understanding of the nature of models and conceptual learning in biology, physics, and chemistry. *International Journal of Science Education, 33*(5), 653-684. http://doi.org/10.1080/09500691003720671

Gogolin, S. & Krüger, D. (2018). Students' understanding of the nature and purpose of models. *Journal of Research in Science Teaching, 55*(9), 1313-1338. https://doi.org/10.1002/tea.21453

Gorsuch, R. L. (1983). Factor analysis (2nd ed.). Lawrence Erlbaum.

Gottheiner, D. M., & Siegel, M. A. (2012). Experienced middle school science teachers' assessment literacy: Investigating knowledge of students' conceptions in genetics and ways to shape instruction. *Journal of Science Teacher Education, 23*(5), 531-557. https://doi.org/10.1007/s10972-012-9278-z

Gotwals, A. W. (2018). Where are we now? Learning progressions and formative assessment. *Applied Measurement in Education, 31*(2), 157-164. https://doi.org/10.1080/08957347.2017.1408626

Grant, M. J., Button, C. M., & Snook, B. (2017). An evaluation of interrater reliability measures on binary tasks using d-Prime. *Applied psychological measurement, 41*(4), 264-276. https://doi.org/10.1177/0146621616684584

Green, S. K., Johnson, R. L., Kim, D. H., & Pope, N. S. (2007). Ethics in classroom assessment practices: Issues and attitudes. *Teaching and Teacher Education, 23*(7), 999-1011. http://doi.org/0.1016/j.tate.2006.04.042

Grossman, P. L. (1990). *The making of a teacher: Teacher knowledge and teacher education*. Teachers College Press, Teachers College, Columbia University.

Grünkorn, J., Upmeier zu Belzen, A. & Krüger, D. (2014). Assessing student's understandings of biological models and their use in science to evaluate theoretical framework. *International Journal of Science Education, 34*(10) 1651-1684. https://doi.org/10.1080/09500693.2013.873155

Guba, E. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. *Educational Communication and Technology, 29*(2), 75-91.

Guy-Gaytán, C., Gouvea, J. S., Griesemer, C., & Passmore, C. (2019). Tensions between learning models and engaging in modeling. *Science & Education, 28*(8), 843-864. https://doi.org/10.1007/s11191-019-00064-y

Gwet, K. (2001). *Handbook of inter-rater reliability: How to estimate the level of agreement between two or multiple raters*. STATAXIS Publishing Company.

Gwet, K. (2008). Computing inter-rater reliability in the presence of high agreement. British *Journal of Mathematical & Statistical Methodology, 61*(1), 29-48. http://doi.org/10.1348/000711006×126600

Hadi, N. U., Abdullah, N., & Sentosa, I. (2016). An easy approach to exploratory factor analysis: Marketing perspective. *Journal of Educational and Social Research, 6*(1), 215-223. http://doi.org/10.5901/jesr.2016.v6n1p215

Hailaya, W., Alagumalai, S., & Ben, F. (2014). Examining the utility of Assessment Literacy Inventory and its portability to education systems in the Asia Pacific region. *Australian Journal of Education, 58*(3), 297-317. https://doi.org/10.1177/0004944114542984

Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher, 20*, 2–7. https://doi.org/10.3102/0013189X020005002

Hall, D. A., Zaragoza Domingo, S., Hamdache, L. Z., Manchaiah, V., Thammaiah, S., Evans, C., Wong, L., & International Collegium of Rehabilitative Audiology and TINnitus Research NETwork. (2018). A good practice guide for translating and adapting hearing-related questionnaires for different languages and cultures. *International Journal of Audiology, 57*(3), 161-175. http/doi.org/10.1080/14992027.2017.1393565

Hammersley, M. (2019). From positivism to post-positivism: Progress or digression? *Teoria Polityki,* (3), 175-188. http://doi.org/10.4467/25440845TP.19.009.10292

Hanrahan, S. J., & Isaacs, G. (2001). Assessing self- and peer assessment: The students' views. *Higher Education Research & Development, 20*(1), 53-70. http://doi.org/10.1080/07294360123776

Harel, O. (2009). The estimation of R^2 and adjusted R^2 in incomplete data sets using multiple imputation. *Journal of Applied Statistics, 36*(10), 1109-1118. https://doi.org/10.1080/02664760802553000

Harlen, W. (2006). *The role of teachers in the assessment of learning*. Newcastle Document Services.

Harrison, A. G., & Treagust, D. F. (2000). Learning about atoms, molecules, and chemical bonds: A case study of multiple-model use in grade 11 chemistry. *Science Education, 84*(3), 352-381. https://doi.org/10.1002/(SICI)1098-237X(200005)84:3%3C352::AID-SCE3%3E3.0.CO;2-J

Hasni, A., Bousadra, F., Belletête, V., Benabdallah, A., Nicole, M. C., & Dumais, N. (2016). Trends in research on project-based science and technology teaching and learning at K–12 levels: a systematic review. *Studies in Science Education, 52*(2), 199-231. https://doi.org/10.1080/03057267.2016.1226573

Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and ltenls. *Applied Psychological Measurement, 9*(2), 139-164. https://doi.org/10.1177/014662168500900204

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112. http://doi.org/10.3102/003465430298487

Hawley, S., Ali, M. S., Berencsi, K., Judge, A., & Prieto-Alhambra, D. (2019). Sample size and power considerations for ordinary least squares interrupted time series analysis: a simulation study. *Clinical Epidemiology, 11*, 197. http://doi.org10.2147/CLEP.S176723

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development, 34*, 177-189. https://doi.org/10.1080/07481756.2002.12069034

Henze, I., van Driel, J. H., & Verloop, N. (2007). Science teachers' knowledge about teaching models and modelling in the context of a new syllabus on public understanding of

science. *Research in Science Education, 37*(2), 99-122. https://doi.org/10.1007/s11165-006-9017-6

Herdman, M., Fox-Rushby, J., & Badia, X. (1998). A model of equivalence in the cultural adaptation of HRQoL instruments: the universalist approach. *Quality of life Research, 7*(4), 323-335. http://doi.org/10.1023/a:1024985930536

Hernández, M. I., Couso, D., & Pintó, R. (2015). Analyzing students' learning progressions throughout a teaching sequence on acoustic properties of materials with a model-based inquiry approach. *Journal of Science Education and Technology, 24*(2-3), 356-377. http://doi.org/10.1007%2Fs10956-014-9503-y

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods, 1*, 104–121. https://doi.org/10.1177/109442819800100106

Hinton, D.,m & Platt, T. (2019). Measurement theory and psychological scaling. In P. M. W. Hackett (Ed.), *Quantitative research methods in consumer psychology* (pp. 59-87). Routledge.

Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management, 24*(5), 623-641. https://doi.org/10.1016/S0149-2063(99)80077-4

Honda, C., & Ohyama, T. (2020). Homogeneity score test of AC 1 statistics and estimation of common AC 1 in multiple or stratified inter-rater agreement studies. *BMC Medical Research Methodology, 20*(1), 1-13. http://doi.org/10.1186/s12874-019-0887-5

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrica, 30*, 179-185. https://doi.org/10.1007/BF02289447

Howard, M. C. (2016). A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human-Computer Interaction, 32*(1), 51-62. https://doi.org/10.1080/10447318.2015.1087664

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research, 103*(2), 219. https://doi.org/10.1007/s11205-011-9843-4

Hutcheson, G. D. (2011). Ordinary least-squares regression. In L. Moutinho & G. D. (Ed.), *The SAGE dictionary of quantitative management research* (pp. 224-228). SAGE.

Ilesanmi, O. O. (2009). What is cross-cultural research. *International Journal of Psychological Studies, 1*(2), 82-96. http://doi.org/10.5539/ijps.v1n2p82

John, M. T., Reißmann, D. R., Feuerstahler, L., Waller, N., Baba, K., Larsson, P., Čelebić, A., Szabo, G., & Rener-Sitar, K. (2014). Exploratory factor analysis of the oral health impact profile. *Journal of Oral Rehabilitation, 41*(9), 635-643. http://doi.org/10.1111/joor.12192

Johnson-Laird, P. N. (1980). Mental models in cognitive science. *Cognitive Science, 4*(1), 71-115. https://doi.org/10.1016/S0364-0213(81)80005-5

Jolliffe, I. T. (1972). Discarding variables in a principal component analysis. I: Artificial data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 21(2), 160-173. https://doi.org/10.2307/2346488

Jonassen, D., & Cho, Y. H. (2008). Externalizing mental models with mindtools. In D. Ifenthaler, P. Pirnay-Dummer & J. M. Spector (Eds.), *Understanding models for learning and instruction* (pp. 145-159). Springer.

Jones, A., & Moreland, J. (2005). The importance of pedagogical content knowledge in assessment for learning practices: A case study of a whole school approach. *The Curriculum Journal, 16* (2), 193–206.

Justi, R. S. (2000). Teaching with historical models. In J. K. Gilbert, & C. J. Boulter (Eds.), *Developing models in science education* (pp. 209-226). Springer.

Justi, R. (2009). Learning how to model in science classroom: Key teacher's role in supporting the development of students' modelling skills. *Educación Química, 20*(1), 32-40. https://doi.org/10.1016/S0187-893X(18)30005-3

Justi, R. & Gilbert, J. K. (2002a). Modelling teachers' views on the nature of modelling, and implications for the education of modellers. *International Journal of Science Education, 24*(4), 369–387. https://doi.org/10.1080/09500690110110142

Justi, R. S., & Gilbert, J. K. (2002b). Science teachers' knowledge about and attitudes towards the use of models and modelling in learning science. *International Journal of science education, 24*(12), 1273-1292. https://doi.org/10.1080/09500690210163198

Justi, R., & Gilbert, J. K. (2003). Teachers' views on the nature of models. *International Journal of Science Education, 25*(11), 1369–1386. https://doi.org/10.1080/0950069032000070324

Justi, R., & Van Driel, J. (2005). The development of science teachers' knowledge on models and modelling: promoting, characterizing, and understanding the process. *International Journal of Science Education, 27*(5), 549-573. https://doi.org/10.1080/0950069042000323773

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151. https://doi.org/10.1177/001316446002000116

Karstad, K., Rugulies, R., Skotte, J., Munch, P. K., Greiner, B. A., Burdorf, A., Søgaard, K., & Holtermann, A. (2018). Inter-rater reliability of direct observations of the physical and psychosocial working conditions in eldercare: An evaluation in the DOSES project. *Applied Ergonomics, 69*, 93-103. http://doi.org/10.1016/j.apergo.2018.01.004

Kaufman, J. D., & Dunlap, W. P. (2000). Determining the number of factors to retain: Q windows-based FORTRAN-IMSL program for parallel analysis. *Behavior Research Methods, Instruments, & Computers, 32*(3), 389-395. https://doi.org/10.3758/BF03200806

Kawasaki, J., & Sandoval, W. A. (2020). Examining teachers' classroom strategies to understand their goals for student learning around the science practices in the Next Generation Science Standards. *Journal of Science Teacher Education, 31*(4), 384-400. https://doi.org/10.1080/1046560X.2019.1709726

Ketokivi, M., & Mantere, S. (2010). Two strategies for inductive reasoning in organizational research. *Academy of Management Review, 35*(2), 315–333. https://doi.org/10.5465/amr.35.2.zok315

Khan, S. (2007). Model-based inquiries in chemistry. *Science Education, 91*, 877–905. https://doi.org/10.1002/sce.20226

Khan, S. (2008b). What if scenarios for testing student models in chemistry. In J. J. Clement, M. & A. Rea-Ramirez (Eds.), *Model based learning and instruction in science* (pp. 139-150). Springer.

Khan, S. (2011a). What's missing in model-based teaching. *Journal of Science Teacher Education, 22*(6), 535-560. http://doi.org/10.1007/s10972-011-9248-x

Khan, S. (2011b). New pedagogies on teaching science with computer simulations. *Journal of Science Education and Technology, 20*(3), 215-232. https://doi.org/10.1007/s10956-010-9247-2

Khan, S. (2012). A hidden GEM: A pedagogical approach to using technology to teach global warming. *The Science Teacher, 79*(8), 59-62.

Khan, S., & VanWynsberghe, R. (2008). Cultivating the under-mined: Knowledge mobilization through cross-case analysis. *Forum: Qualitative Social Research, 9*, 1–21. https://doi.org/10.17169/fqs-9.1.334

Kind, P., Jones, K., & Barmby, P. (2007). Developing attitudes towards science measures. *International Journal of Science Education, 29*(7), 871-893. https://doi.org/10.1080/09500690600909091

Kind, V. (2009). Pedagogical content knowledge in science education: perspectives and potential for progress. *Studies in Science Education, 45*(2), 169-204. https://doi.org/10.1080/03057260903142285

King, N. (2004). Using templates in the thematic analysis of text. In C. Cassell & G. Symon (Eds.), *Essential guide to qualitative methods in organizational research* (pp. 257–270). Sage.

Kivunja, C., & Kuyini, A. B. (2017). Understanding and applying research paradigms in educational contexts. *International Journal of Higher Education, 6*(5), 26-41. https://doi.org/10.5430/ijhe.v6n5p26

Kline, R. B. (2011). Principles and practice of structural equation modeling (3rd ed.). Guilford.

Kline, P. (2014). An easy guide to factor analysis. Routledge.

Koh, K. H. (2011). Improving teachers' assessment literacy through professional development. *Teaching Education, 22*(3), 255-276. https://doi.org/10.1080/10476210.2011.593164

Korstjens, I., & Moser, A. (2018). Series: Practical guidance to qualitative research. Part 4: trustworthiness and publishing. *European Journal of General Practice, 24*(1), 120-124. https://doi.org/10.1080/13814788.2017.1375092

Krajcik, J., & Merritt, J. (2012). Engaging students in scientific practices: What does constructing and revising models look like in the science classroom? *The Science Teacher, 79*(3), 6-10.

Krell, M., Reinisch, B., & Krüger, D. (2015). Analyzing students' understanding of models and modeling referring to the disciplines biology, chemistry, and physics. *Research in Science Education, 45*(3), 367-393. https://doi.org/10.1007/s11165-014-9427-9

Krell, M., zu Belzen, A. U., & Krüger, D. (2014). Students' levels of understanding models and modelling in biology: Global or aspect-dependent? *Research in science education, 44*(1), 109-132. https://doi.org/10.1007/s11165-013-9365-y

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Kumtepe, E. G., Kumtepe, A., Uğurhan Y. (2019). Investigating the new media literacy skills of open and distance leaners. In C. Erdem, H. Bağci, & M. Koçyiğit (Eds.), *21st century skills and education* (pp. 112-137). Cambridge Scholars Publishing.

Larson, R. B. (2019). Controlling social desirability bias. *International Journal of Market Research, 61*(5), 534-547. https://doi.org/10.1177/1470785318805305

Lattery, M. (2017). *Deep learning in introductory Physics: Exploratory studies of model-based reasoning*. The United States of America: Information Age Publishing.

Lautenschlager,G. J. (1989). A comparison of alternatives to conducting Monte Carlo analyses

for determining parallel analysis criteria. *Multivariate Behavioral Research, 24*, 365-395. https://doi.org/10.1207/s15327906mbr2403_6

Lehrer, R., & Schauble, L. (2010). What kind of explanation is a model? In M. K. Stein & L. Kucan (Eds.), *Instructional explanations in the disciplines* (pp. 9-22). Springer.

Levers, M. J. D. (2013). Philosophical paradigms, grounded theory, and perspectives on emergence. *Sage Open, 3*(4), 1-6. https://doi.org/10.1177/2158244013517243

Levy-Vered, A., & Nasser-Abu Alhija, F. (2015). Modelling beginning teachers' assessment literacy: the contribution of training, self-efficacy, and conceptions of assessment. *Educational Research and Evaluation, 21*(5-6), 378-406. https://doi.org/10.1080/13803611.2015.1117980

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., Clarke, M., Kleijnen, J, & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Medicine, 6*(7), 1-28. https://doi.org/10.1136/bmj.b2700

Lin, J. W., & Chiu, M. H. (2010). The mismatch between students' mental models of acids/bases and their sources and their teacher's anticipations thereof. *International Journal of Science Education, 32*(12), 1617-1646. https://doi.org/10.1080/09500690903173643

Lincoln, Y. S., & Guba, E. G. (1986). But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation. *New Directions for Program Evaluation, 1986*(30), 73- 84. http://doi.org/10.1002/ev.1427

Liu, S. C. (2005). Models of "the heavens and the earth": An investigation of German and Taiwanese students' alternative conceptions of the universe. *International Journal of Science and Mathematics Education, 3*(2), 295-325.

Liu, S. C., & Lin, H. S. (2015). Exploring undergraduate students' mental models of the environment: Are they related to environmental affect and behavior? *The Journal of Environmental Education*, *46*(1), 23-40. http://doi.org/10.1080/00958964.2014.953021

Liu,Y., Zumbo, B. D., & Wu, A. D. (2014). Relative Importance of Predictors in Multilevel Modeling, *Journal of Modern Applied Statistical Methods, 13*(1), 2-22. http://doi.org/10.22237/jmasm/1398916860

Luxford, C. J., & Bretz, S. L. (2013). Moving beyond definitions: what student-generated models reveal about their understanding of covalent bonding and ionic bonding. *Chemistry Education Research and Practice*, *14*(2), 214-222. http://doi.org/10.1039/c3rp20154f

Lynn, M.R. (1986). Determination and quantification of content validity. *Nursing Research, 35*, 382– 385.

Magnusson, S., Krajcik, J., & Borko, H. (1999). Nature, sources, and development of pedagogical content knowledge for science teaching. In J. Gess-Newsome & N. G. Lederman (Eds.), *Examining pedagogical content knowledge* (pp. 95–132). Kluwer.

Mahtani, K. R., Heneghan, C., & Aronson, J. (2020). Single screening or double screening for study selection in systematic reviews? *BMJ evidence-based medicine*, *25*(4), 1-2. http://dx.doi.org/10.1136/bmjebm-2019-111269

Mak, P. (2019). Impact of professional development programme on teachers' competencies in assessment. *Journal of Education for Teaching, 45*(4), 481-485. https://doi.org/10.1080/02607476.2019.1639266

Margot, K. C., & Kettler, T. (2019). Teachers' perception of STEM integration and education: A systematic literature review. *International Journal of STEM Education, 6(*1), 1-16. https://doi.org/10.1186/s40594-018-0151-2

Markus K. A., & Lin C. (2012). Construct validity. In N. J. Salkind, *Encyclopedia of research design* (pp. 230-233). Sage. http://doi.org/10.4135/9781412961288

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*(3), 320-341. https://doi.org/10.1207/s15328007sem1103_2

Martin, A. M., & Hand, B. (2009). Factors affecting the implementation of argument in the elementary science classroom. A longitudinal case study. *Research in Science Education, 39*(1), 17-38. https://doi.org/10.1007/s11165-007-9072-7

Martínez, J. F., Stecher, B., & Borko, H. (2009). Classroom assessment practices, teacher judgments, and student achievement in mathematics: Evidence from the ECLS. *Educational Assessment, 14*(2), 78-102. http://doi.org/10.1080/10627190903039429

Mayer, K., & Krajcik, J. (2015). Designing and assessing scientific modeling tasks. In R. Gunstone (Eds.), *Encyclopedia of science education* (pp. 291-297). Springer.

McChesney, K., & Aldridge, J. (2019). Weaving an interpretivist stance throughout mixed methods research. *International Journal of Research & Method in Education, 42*(3), 225-238. https://doi.org/10.1080/1743727X.2019.1590811

McGregor, S. L., & Murnane, J. A. (2010). Paradigm, methodology and method: Intellectual integrity in consumer scholarship. *International Journal of Consumer Studies*, 34(4), 419-427. https://doi.org/10.1111/j.1470-6431.2010.00883.x

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica, 22*(3), 276-282.

McManus, S. (2008). *Attributes of effective formative assessment*. Washington, DC: Council of Chief State School Officers. Retrieved from https://ccsso.org/sites/default/files/2017-12/Attributes_of_Effective_2008.pdf

Mellati, M., & Khademi, M. (2018). Exploring teachers' assessment literacy: Impact on learners' writing achievements and implications for teacher development. *Australian Journal of Teacher Education, 43*(6), 1. http://doi.org/10.14221/AJTE.2018V43N6.1

Merritt, J., & Krajcik, J. (2013). Learning progression developed to support students in building a particle model of matter. In G. Tsaparlis & H. Sevian (Eds.), *Concepts of matter in science education* (pp. 11-45). Springer.

Mertler, C. A. (2004). Secondary teachers' assessment literacy: Does classroom experience make a difference? *American Secondary Education, 33*(1), 49-64.

Mertler, C. A., & Campbell, C. (2005). *Measuring teachers' knowledge & application of classroom assessment concepts: Development of the assessment literacy inventory*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada, 11–15 April 2005.

Mertler, C. A., & Reinhart, R. V. (2016). *Advanced and multivariate statistical methods: Practical application and interpretation*. Routledge.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement, 21*, 215-237.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Sage.

MINEDUC. (2015). *Biología. Programa de estudio: Cuarto año medio, Actualización 2009* [Biology: 12th grade study plan, update 2009]. Ministerio de Educación de Chile.

MINEDUC. (2019). *Bases Curriculares 3º y 4º Medio*. [Curricular Bases 11th and 12th grade]. Ministerio de Educación de Chile.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews, 4*(1), 1. https://doi.org/10.1186/2046-4053-4-1

Moran, J., Alexander, V. D., Cronin, A., Dickinson, M., Fielding, J., Sleney, J., & Thomas, H. (2006). Triangulation and integration: processes, claims and implications. *Qualitative Research, 6*(1), 45-59. http://doi.org/10.1177/1468794106058870

Morrow, S. L. (2005). Quality and trustworthiness in qualitative research in counseling psychology. *Journal of Counseling Psychology, 52*(2), 250-260. https://doi.org/10.1037/0022-0167.52.2.250

Mruck, K., & Breuer, F. (2003). Subjectivity and reflexivity in qualitative research—A new FQS Issue. *Historical Social Research/Historische Sozialforschung*, *28*(3), 189-212.

Muhammad, N. B., Ali, N. B. M., Zamani, S. B., Yamin, N. A. B., & Ismail, N. N. B. (2020). Examining assessment literacy: A study of technical teacher. *European Journal of Molecular & Clinical Medicine, 7*(8), 705-717.

Multon, K. D., & Coleman, J. (2018). Inter-rater reliability. In B. B. Frey (Ed.), *The SAGE encyclopedia of educational research, measurement, and evaluation* (pp. 863-865). Sage Publications.

Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing, 5*(2), 159–168. https://doi.org/10.1207/s15327574ijt0502_4

Munns, G., & Woodward, H. (2006). Student engagement and student self-assessment: the REAL framework. *Assessment in Education: Principles, Policy & Practice, 13*(2), 193-213. https://doi.org/10.1080/09695940600703969

Namdar, B., & Shen, J. (2015). Modeling-oriented assessment in K-12 science education: A synthesis of research from 1980 to 2013 and new directions. *International Journal of Science Education, 37*(7), 993-1023. https://doi.org/10.1080/09500693.2015.1012185

Nayak, B. K. (2010). Understanding the relevance of sample size calculation. *Indian Journal of Ophthalmology, 58*(6), 469. http://doi.org/10.4103/0301-4738.71673

Nersessian, N. J. (2013). Mental modeling in conceptual change. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 395–411). New York, NY: Routledge

Neuman, L. W (2007). The meanings of methodology. In L. W. Neuman (Ed.), *Social research methods: Qualitative and quantitative approaches* (pp. 91-124). Allyn and Bacon.

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. Retrieved from http://www.nextgenscience.org/

Noble, H., & Smith, J. (2015). Issues of validity and reliability in qualitative research. *Evidence-based Nursing, 18*(2), 34-35. http://dx.doi.org/10.1136/eb-2015-102054

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods, 16*(1),1-13. http://doi.org/1609406917733847

Nunez-Oviedo, M. C., & Clement, J. (2008). A competition strategy and other discussion modes for developing mental models in large group discussion. In J. Clement & M. A. Rea-Ramirez (Eds.), *Model based learning and instruction in science* (pp. 117–138). Springer.

Nunez-Oviedo, M. C., & Clement, J. J. (2019). Large scale scientific modeling practices that can organize science instruction at the unit and lesson levels. *Frontiers in Education, 4*(68), 1-22. https://doi.org/10.3389/feduc.2019.00068

Nunnally, J. C., & Bernstein, I. R. (1994). *Psychometric theory* (3rd Edition). McGraw-Hill.

OECD. (2003). *Attracting, developing and retaining effective teachers: OECD Activity: Country background report for Chile*. OECD publishing.

OECD. (2011). *Education at a Glance: OECD indicators*. OECD Publishing.

OECD. (2017). *PISA 2015 Results (Volume I): Excellence and equity in education*. OECD publishing.

OECD. (2019a). PISA 2018 Results (Volume I): What students know and can do. OECD publishing.

OECD. (2019b). *TALIS (Volume I): Teachers and school leaders as lifelong learners*. OECD publishing.

Ogan-Bekiroglu, F. (2007). Effects of model-based teaching on preservice physics teachers' conceptions of the moon, moon phases, and other lunar phenomena. *International Journal of Science Education, 29*, 555–593. https://doi.org/10.1080/09500690600718104

Ogan-Bekiroglu, F., & Suzuk, E. (2014). Pre-service teachers' assessment literacy and its implementation into practice. *Curriculum Journal, 25*(3), 344-371. https://doi.org/10.1080/09585176.2014.899916

Oh, P. S., & Oh, S. J. (2011). What teachers of science need to know about models: An overview. *International Journal of Science Education, 33*(8), 1109-1130. https://doi.org/10.1080/09500693.2010.502191

Oliva, J. M., & Blanco-López, A. (2021). Development of a questionnaire for assessing Spanish-speaking students' understanding of the nature of models and their uses in science. *Journal of Research in Science Teaching, 58*(6), (852-878). https://doi.org/10.1002/tea.21681

Osborne, J. (2014). Best practices in exploratory factor analysis. CreateSpace Independent Publishing.

Otero, V. K. (2006). Moving beyond the ‘‘Get it or don’t’’ conception of formative assessment. *Journal of Teacher Education, 57*, 247–255. https://doi.org/10.1177/0022487105285963

Özcan, Ö. (2015). Investigating students’ mental models about the nature of light in different contexts. *European Journal of Physics, 36*(6), 1-16. http://doi.org/10.1088/0143-0807/36/6/065042

Palaganas, E. C., Sanchez, M. C., Molintas, V. P., & Caricativo, R. D. (2017). Reflexivity in qualitative research: A journey of learning. *Qualitative Report, 22*(2), 426-436. https://doi.org/10.46743/2160-3715/2017.2552

Pang, M., Ho, T. M., & Man, R. (2009). Learning approaches and outcome-based teaching and learning: A case study in Hong Kong, China. *Journal of Teaching in International Business, 20*(2), 106-122. http://doi.org/10.1080/08975930902827825

Park, S., & Oliver, J. S. (2008). Revisiting the conceptualisation of pedagogical content knowledge (PCK): PCK as a conceptual tool to understand teachers as professionals. *Research in Science Education, 38*(3), 261-284. https://doi.org/10.1007/s11165-007-9049-6

Parr, J. M., & Timperley, H. S. (2008). Teachers, schools and using evidence: Considerations of preparedness. *Assessment in Education: Principles, Policy & Practice, 15*(1), 57-71. https://doi.org/10.1080/09695940701876151

Passmore, C., Gouvea, J. S., & Giere, R. (2014). Models in science and in learning science: Focusing scientific practice on sense-making. In M. R. Matthews (Ed.), *International handbook of research in history, philosophy and science teaching* (pp. 1171-1202). Springer.

Patron, E., Wikman, S., Edfors, I., Johansson-Cederblad, B., & Linder, C. (2017). Teachers' reasoning: Classroom visual representational practices in the context of introductory chemical bonding. *Science Education, 101*(6), 887-906.

Pek, J., Wong, O., & Wong, A. (2018). How to address non-normality: A taxonomy of approaches, reviewed, and illustrated. *Frontiers in Psychology, 9*, 2104. https://doi.org/10.3389/fpsyg.2018.02104

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Piedmont R.L. (2014). Construct Validity. In A. C. Michalos (Ed.), *Encyclopedia of quality of life and well-being research* (pp. 1212-1212). Springer.

Pierson, A. E., Clark, D. B., & Sherard, M. K. (2017). Learning progressions in context: Tensions and insights from a semester-long middle school modeling curriculum. *Science Education, 101*(6), 1061–1088. https://doi.org/10.1002/sce.21314

Plake, B. S. (1993). Teacher assessment literacy: Teachers' competencies in the educational assessment of students. *Mid-Western Educational Researcher, 6*(1), 21–27.

Plake, B. S., Impara, J. C., & Fager, J. J. (1993). Assessment competencies of teachers: A national survey. *Educational Measurement: Issues and Practice, 12*(4), 10-12.

Pluta, W. J., Chinn, C. A., & Duncan, R. G. (2011). Learners' epistemic criteria for good scientific models. *Journal of Research in Science Teaching, 48*(5), 486-511. https://doi.org/10.1002/tea.20415

Polit, D., & Beck, C. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health, 29*, 489-497. http://doi.org/10.1002/nur.20147

Ponterotto, J. G. (2005). Qualitative research in counselling psychology: A primer on research paradigms and philosophy of science. *Journal of Counseling Psychology, 55*, 126–136. https://doi.org/10.1037/0022-0167.52.2.126

Pope, N., Green, S. K., Johnson, R. L., & Mitchell, M. (2009). Examining teacher ethical dilemmas in classroom assessment. *Teaching and Teacher Education, 25*(5), 778-782. http://doi.org/10.1016/j.tate.2008.11.013

Popham, W. J. (1991). Appropriateness of teachers' test-preparation practices. *Educational Measurement: Issues and Practice, 10*(4), 12–15.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory into Practice, 48*(1), 4-11. https://doi.org/10.1080/00405840802577536

Potvin, P., & Hasni, A. (2014). Interest, motivation and attitude towards science and technology at K-12 levels: a systematic review of 12 years of educational research. *Studies in Science Education, 50*(1), 85-129. https://doi.org/10.1080/03057267.2014.881626

Prenger, R., & Schildkamp, K. (2018). Data-based decision making for teacher and student learning: a psychological perspective on the role of the teacher. *Educational Psychology,* 1-19. https://doi.org/10.1080/01443410.2018.1426834

Quigley, C., Oliviera, A. W., Curry, A., & Buck, G. (2011). Issues and techniques in translating scientific terms from English to Khmer for a university-level text in Cambodia. *Language, Culture and Curriculum, 24*(2), 159-177. https://doi.org/10.1080/07908318.2011.583663

Quillin, K., & Thomas, S. (2015). Drawing-to-learn: a framework for using drawings to promote model-based reasoning in biology. *CBE-Life Sciences Education, 14*(1), 1-16. http://doi.org/10.1187/cbe.14-08-0128

Raghavan, K., Sartoris, M. L., & Glaser, R. (1998). Why does it go up? The impact of the MARS curriculum as revealed through changes in student explanations of a helium balloon. *Journal of Research in Science Teaching, 35*(5), 547-567. https://doi.org/10.1002/(SICI)1098-2736(199805)35:5%3C547::AID-TEA5%3E3.0.CO;2-P

Ravanal, M., López-Cortés, F. y Rodríguez, M. (2018). Creencias de profesores chilenos de biología sobre la preparación de la enseñanza [Belief of Chilean biology teacher on the preparation of teaching]. *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias, 15*(3), 3601-3616.

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., Thissen, D., Revicki, D., Weiss, D., Hambleton, R., Liu, H., Gershon, R., Reise, S., Lai, J., Cella, D., & PRIMIS Cooperative Group. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes

Measurement Information System (PROMIS). *Medical Care, 45*(5), 22-31. http://doi.org/10.1097/01.mlr.0000250483.85507.04

Reinholz, D. (2016). The assessment cycle: a model for learning through peer assessment. *Assessment & Evaluation in Higher Education, 41*(2), 301-315. https://doi.org/10.1080/02602938.2015.1008982

Remesal, A. (2007). Educational reform and primary and secondary teachers' conceptions of assessment: The Spanish instance, building upon Black and Wiliam (2005). *The Curriculum Journal, 18*, 27–38. https://doi.org/10.1080/09585170701292133

Rinehart, R., Duncan, R., Chinn, C., Atkins, T., & DiBenedetti, J. (2016). Critical design decisions for successful model-based inquiry in science classrooms. *International Journal of Designs for Learning, 7*(2).

Rodríguez Amador, R., & López Yáñez, J. (2020). Creencias y prácticas curriculares de docentes chilenos de Física en Educación Secundaria [Beliefs and curricular practices of Chilean physics teachers in secondary education]. *Enseñanza de las Ciencias, 38*(2), 121-139.

Rogers, F., Huddle, P. A., & White, M. D. (2000). Using a teaching model to correct known misconceptions in electrochemistry. *Journal of chemical education, 77*(1), 104. https://doi.org/10.1021/ed077p104

Ruiz-Primo, M. A., & Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *Journal of Research in Science Teaching, 44*, 57–84. https://doi.org/10.1002/tea.20163

Ruppert, J., Duncan, R. G., & Chinn, C. A. (2017). Disentangling the role of domain-specific knowledge in student modeling. *Research in Science Education*, 1-28. https://10.1007/s11165-017-9656-9

Ryan, C., Hesselgreaves, H., Wu, O., Paul, J., Dixon-Hughes, J., & Moss, J. G. (2018). Protocol for a systematic review and thematic synthesis of patient experiences of central venous access devices in anti-cancer treatment. *Systematic Reviews, 7*(1), 61. https://doi.org/10.1186/s13643-018-0721-x

Sayers, A. (2008). Tips and tricks in performing a systematic review. *British Journal of General Practice, 58*(547), 136-136.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical methods in medical research, 8*(1), 3-15. http://doi.org/10.1177/096228029900800102

Schneider, R. M., & Plasman, K. (2011). Science teacher learning progressions: A review of science teachers' pedagogical content knowledge development. *Review of Educational Research, 81*(4), 530-565. https://doi.org/10.3102/0034654311423382

Schafer, W. D. (1993). Assessment literacy for teachers. *Theory into Practice, 32*(2), 118-126. https://doi.org/10.1080/00405849309543585

Scharp, K. M., & Sanders, M. L. (2019). What is a theme? Teaching thematic analysis in qualitative communication research methods. Communication Teacher, 33(2), 117-121. http://doi.org/10.1080/17404622.2018.1536794

Schwandt, T. A. (1998). Constructivist, Interpretivist approaches to human inquiry. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Landscape of Qualitative Research: Theories and Issues* (pp. 221–259). Sage.

Schwartz, P., & Barbera, J. (2014). Evaluating the content and response process validity of data from the chemical concepts inventory. *Journal of Chemical Education, 91*(5), 630-640. https://doi.org/10.1021/ed400716p

Schwartz, D. L., & Hartman, K. (2007). It's not television anymore: Designing digital video for learning and assessment. In R. Goldman, R. Pea, B. Barron, & S. J. Derry (Eds.), *Video research in the learning sciences* (pp. 335–348). Erlbaum.

Schwarz, C., Reiser, B. J., Acher, A., Kenyon, L., & Fortus, D. (2012). MoDeLS. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progression in science: Current challenges and future directions* (pp. 101-137). Sense Publishers.

Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Achér, A., Fortus, D., Shwartz, Y., & Krajcik, J. (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *Journal of Research in Science Teaching, 46*(6), 632-654. https://doi.org/10.1002/tea.20311

Schwarz, C. V., & White, B. Y. (2005). Metamodeling knowledge: Developing students' understanding of scientific modeling. *Cognition and Instruction, 23*(2), 165-205. https://doi.org/10.1207/s1532690xci2302_1

Scott, G. W. (2017). Active engagement with assessment and feedback can improve group-work outcomes and boost student confidence. *Higher Education Pedagogies, 2*(1), 1-13. https://doi.org/10.1080/23752696.2017.1307692

Senocak, E., & Baloglu, M. (2014). The adaptation and preliminary psychometric properties of the Derived Chemistry Anxiety Rating Scale. *Chemistry Education Research and Practice, 15*(4), 800-806. http://doi.org/0.1039/c4rp00073k

Serdar, C. C., Cihan, M., Yücel, D., & Serdar, M. A. (2021). Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. *Biochemia Medica, 31*(1), 27-53. http://doi.org/10.11613/BM.2021.010502

Shenton, A. K. (2004). Strategies for ensuring trustworthiness in qualitative research projects. *Education for Information, 22*(2), 63-75. http://doi.org/10.3233/EFI-2004-22201

Shepard, L. A. (2018). Learning progressions as tools for assessment and learning. *Applied Measurement in Education, 31*(2), 165-174. https://doi.org/10.1080/08957347.2017.1408628

Shepardson, D. P., Wee, B., Priddy, M., & Harbor, J. (2007). Students' mental models of the environment. *Journal of Research in Science Teaching, 44*(2), 327-348. http://dx.doi.org/10.1002/tea.20161

Shi, D., Lee, T., & Maydeu-Olivares, A. (2019). Understanding the model size effect on SEM fit indices. *Educational and Psychological Measurement, 79*(2), 310-334. https://doi.org/10.1177/0013164418783530

Shrotryia, V. K., & Dhanda, U. (2019). Content validity of assessment instrument for employee engagement. *SAGE Open, 9*(1), 1-7.

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4-14. https://doi.org/10.3102/0013189X015002004

Siegel, M. A., & Wissehr, C. (2011). Preparing for the plunge: Preservice teachers' assessment literacy. *Journal of Science Teacher Education, 22*(4), 371-391. https://doi.org/10.1007/s10972-011-9231-6

Siegel, M. A., Wissehr, C., & Halverson, K. (2008). Sounds like success: A framework for equitable assessment. *The Science Teacher, 75*(3), 43–46.

Siweya, H. J., & Letsoalo, P. (2014). Formative assessment by first-year chemistry students as predictor of success in summative assessment at a South African university. *Chemistry Education Research and Practice, 15*(4), 541-549. https://doi.org/10.1039/C4RP00032C

Smith, J. A. (Ed.). (2004). *Qualitative psychology: A practical guide to methods*. SAGE.

Smith, K. C., & Alonso, V. (2020). Measuring student engagement in the undergraduate general chemistry laboratory. *Chemistry Education Research and Practice*. http://doi.org/10.1039/c8rp00167g

Smith, R. J., Lehning, A. J., & Dunkle, R. E. (2013). Conceptualizing age-friendly community characteristics in a sample of urban elders: an exploratory factor analysis. *Journal of Gerontological Social Work, 56*(2), 90-111. http://doi.org/10.1080/01634372.2012.739267

Stewart, J., Cartier, J. L., & Passmore, C. M. (2005). Developing understanding through model-based inquiry. In M. S. Donovan & J. D. Brandsford (Eds.), *How students learn: Science in the classroom* (pp. 515-565). The National Academic Press.

Stiggins, R. J. (1991a). Relevant classroom assessment training for teachers. *Educational Measurement: Issues and Practice, 10*(1), 7-12. https://doi.org/10.1111/j.1745-3992.1991.tb00171.x

Stiggins, R. J. (1991b). Assessment literacy. Phi Delta Kappan, 72, 534–539.

Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan, 77*(3), 238–245.

Stiggins, R. J., & Conklin, N. F. (1992). *In teachers' hands: Investigating the practices of classroom assessment*. State University of New York Press.

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (2nd ed.). SAGE.

Strickland, J. (2017). *Logistic regression inside-out. With R studio and SAS studio*. Lulu

Sunyono, L., Yuanita, L., & Ibrahim, M. (2015). Supporting students in learning with multiple representation to improve student mental models on atomic structure concepts. *Science Education International, 26*(2), 104-125.

Sutton, J., & Austin, Z. (2015). Qualitative research: Data collection, analysis, and management. *The Canadian Journal of Hospital Pharmacy, 68*(3), 226. http://doi.org/10.4212/cjhp.v68i3.1456

Svoboda, J., & Passmore, C. (2013). The strategies of modeling in biology education. *Science & Education, 22*(1), 119-142. https://doi.org/10.1007/s11191-011-9425-5

Syed, M., & Nelson, S. C. (2015). Guidelines for establishing reliability when coding narrative data. *Emerging Adulthood, 3*(6), 375-387. https://doi.org/10.1177/2167696815587648

Tabachnick, B. G., & Fidel, L. S. (2001). *Using Multivariate Statistics*. Allyn & Bacon.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics (5th ed.)*. Allyn & Bacon.

Taber, K. S. (2000). Case studies and generalizability: Grounded theory and research in science education. *International Journal of Science Education, 22*(5), 469-487. https://doi.org/10.1080/095006900289732

Taber, K. S. (2018a). Lost and found in translation: guidelines for reporting research data in an 'other' language. *Chemistry Education Research and Practice, 19*(3), 646-652. https://doi.org/10.1039/C8RP90006J

Taber, K. S. (2018b). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education, 48*(6), 1273-1296. https://doi.org/10.1007/s11165-016-9602-2

Tacoshi, M., & Fernández, C. (2014). Knowledge of assessment: An important component in the PCK of chemistry teachers. *Problems of Education in the 21st Century, 62*, 124-127.

Teddlie, C., & Yu, F. (2007). Mixed methods sampling: A typology with examples. *Journal of Mixed Methods Research, 1*(1), 77-100. https://doi.org/10.1177/1558689806292430

Thompson B., & Daniel L. G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement, 56*(2), 197–208. https://doi.org/10.1177/0013164496056002001

Tittle, C. K. (1994). Toward an educational psychology of assessment for teaching and learning: Theories, contexts, and validation arguments. *Educational Psychologist, 29*(3), 149–162. https://doi.org/10.1207/s15326985ep2903_4

Ültanir, E. (2012). An epistemologic glance at the constructivist approach: constructivist learning in Dewey, Piaget, and Montessori. *International Journal of Instruction, 5*(2), 195-212.

Upmeier zu Belzen, A., & Krüger,, D. (2010). Modellkompetenz im Biologieunterricht [Model competence in biology education]. *Zeitschrift fu ¨r Didaktik der Naturwissenschaften* [Journal of Science Education], 16, 41–57.

Usry, J., Partington, S. W., & Partington, J. W. (2018). Using expert panels to examine the content validity and inter-rater reliability of the ABLLS-R. *Journal of Developmental and Physical Disabilities, 30*(1), 27-38. http://doi.org/10.1007/S10882-017-9574-9

Vaismoradi, M., Turunen, H., & Bondas, T. (2013). Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study. *Nursing & Health Sciences, 15*(3), 398-405. http://doi.org/10.1111/nhs.12048

Van Driel, J., & Verloop, N. (1998). Pedagogical content knowledge: A unifying element in the knowledge base of teachers. *Pedagogische Studiën, 75,* 225-237.

Van Driel, J., & Verloop, N. (1999). Teachers' knowledge of models and modelling in science. *International Journal of Science Education, 21*(11), 1141-1153. https://doi.org/10.1080/095006999290110

VanWynsberghe, R., & Khan, S. (2007). Redefining case study. *International Journal of Qualitative Methods, 6*(2), 80-94. https://doi.org/10.1177/160940690700600208

Vishnumolakala, V. R., Southam, D. C., Treagust, D. F., & Mocerino, M. (2016). Latent constructs of the students' assessment of their learning gains instrument following instruction in stereochemistry. *Chemistry Education Research and Practice, 17*(2), 309-319. https://doi.org/10.1039/C5RP00214A

Vlachou, M. A. (2018). Classroom assessment practices in middle school science lessons: A study among Greek science teachers. *Cogent Education, 5*(1), 1-9. https://doi.org/10.1080/2331186X.2018.1455633

Vojíř, K., & Rusek, M. (2019). Science education textbook research trends: A systematic literature review. *International Journal of Science Education, 41*(11), 1496-1516. https://doi.org/10.1080/09500693.2019.1613584

Vonderwell, S. & Turner, S. (2005). Active learning and preservice teachers' experiences in an online course: A case study. *Journal of Technology and Teacher Education, 13*(1), 65-84.

Vosniadou, S. (1994). Capturing and modelling the process of conceptual change. *Learning and Instruction, 4*, 45–69. https://doi.org/10.1016/0959-4752(94)90018-3

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive psychology, 24*(4), 535-585. https://doi.org/10.1016/0010-0285(92)90018-W

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Waffenschmidt, S., Knelangen, M., Sieben, W., Bühn, S., & Pieper, D. (2019). Single screening versus conventional double screening for study selection in systematic reviews: a methodological systematic review. *BMC Medical Research Methodology, 19*(1), 132. https://doi.org/10.1186/s12874-019-0782-0

Wakefield, C., Adie, J., Pitt, E., & Owens, T. (2014). Feeding forward from summative assessment: The essay feedback checklist as a learning tool. *Assessment & Evaluation in Higher Education, 39*(2), 253-262. https://doi.org/10.1080/02602938.2013.822845

Wanner, T., & Palmer, E. (2018). Formative self-and peer assessment for improved student learning: the crucial factors of design, teacher participation and feedback. *Assessment & Evaluation in Higher Education, 43*(7), 1032-1047. https://doi.org/10.1080/02602938.2018.1427698

Warrens, M. J. (2010). A formal proof of a paradox associated with Cohen's kappa. *Journal of Classification, 27*(3), 322-332. https://doi.org/10.1007/s00357-010-9060-x

Weed, M. (2009). Research quality considerations for grounded theory research in sport & exercise psychology. *Psychology of Sport and Exercise, 10*(5), 502-510. https://doi.org/10.1016/j.psychsport.2009.02.007

Willey, K., & Gardner, A. (2010). Investigating the capacity of self and peer assessment activities to engage students and promote learning. *European Journal of Engineering Education, 35*(4), 429-443. https://doi.org/10.1080/03043797.2010.490577

Williams, G., & Clement, J. (2015). Identifying multiple levels of discussion-based teaching strategies for constructing scientific models. *International Journal of Science Education, 37*(1), 82-107. https://doi.org/10.1080/09500693.2014.966257

Williams, B., Onsman, A., & Brown, T. (2010). Exploratory factor analysis: A five-step guide for novices. *Australasian Journal of Paramedicine, 8*(3), 1-13.

Willis, J. W. (2007). *Foundations of Qualitative Research*. Sage.

Windschitl, M. (2004). *What types of knowledge do teachers use to engage learners in "doing science?"*. Paper commissioned by the National Academy of Sciences. Washington, DC: Board of Science Education. Retrieved from https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_073331.pdf

Windschitl, M., Thompson, J., & Braaten, M. (2008). How novice science teachers appropriate epistemic discourses around model-based inquiry for use in classrooms. *Cognition and Instruction, 26*(3), 310-378. https://doi.org/10.1080/07370000802177193

Wolak, M. E., Fairbairn, D. J., & Paulsen, Y. R. (2012). Guidelines for estimating repeatability. *Methods in Ecology and Evolution, 3*, 129–137. https://doi.org/10.1111/j.2041-210X.2011.00125.x

Wolfswinkel, J. F., Furtmueller, E., & Wilderom, C. P. (2013). Using grounded theory as a method for rigorously reviewing literature. *European Journal of Information Systems, 22*(1), 45-55. https://doi.org/10.1057/ejis.2011.51

Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K. L. (2013). A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a

study conducted with personality disorder samples. *BMC Medical Research Methodology, 13*(1), 61. http://doi.org/10.1186/1471-2288-13-61

Xu, Y., & Brown, G. T. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education, 58*, 149-162. http://doi.org/10.1016/j.tate.2016.05.010

Xu, W., & Zammit, K. (2020). Applying thematic analysis to education: A hybrid approach to interpreting data in Practitioner research. *International Journal of Qualitative Methods,* 19, 1-9. https://doi.org/10.1177/1609406920918810

Yin, R. K. (2009). *Case study research: Design and methods*. (4th Ed.). SAGE.

Yin, R. K. (2011). *Qualitative research from start to finish*. The Guilford Press.

Yip, D. (2001). Promoting the development of a conceptual change model of science instruction in prospective secondary biology teachers. *International Journal of Science Education, 23*, 755–770. https://doi.org/10.1080/09500690010016067

Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology, 9*(2), 79-94. http://doi.org/10.20982/TQMP.09.2.P079

Yoo, W., Mayberry, R., Bae, S., Singh, K., He, Q. P., & Lillard Jr, J. W. (2014). A study of effects of multicollinearity in the multivariable analysis. *International Journal of Applied Science and Technology*, *4*(5), 9-19.

Yorke, M. (2001). Formative assessment and its relevance to retention. *Higher Education Research & Development, 20*(2), 115-126. https://doi.org/10.1080/758483462

Yu, C., & He, Z.-C. (2017). Analysing the spatial-temporal characteristics of bus travel demand using the heat map. *Journal of Transport Geography, 58*, 247–255. https://doi.org/10.1016/j.jtrangeo.2016.11.009

Zamawe, F. C. (2015). The implication of using NVivo software in qualitative data analysis: Evidence-based reflections. *Malawi Medical Journal, 27*(1), 13-15.

Zangori, L., Vo, T., Forbes, C. T., & Schwarz, C. V. (2017). Supporting 3rd-grade students model-based explanations about groundwater: A quasi-experimental study of a curricular intervention. *International Journal of Science Education, 39*(11), 1421–1442. https://doi.org/10.1080/09500693.2017.1336683

Zec, S., Soriani, N., Comoretto, R., & Baldi, I. (2017). High agreement and high prevalence: the paradox of Cohen's Kappa. *The Open Nursing Journal, 11*, 211. http://doi.org/10.2174/1874434601711010211

Zhang, Z. (2016). Residuals and regression diagnostics: focusing on logistic regression. *Annals of Translational Medicine, 4*(10). http://doi.org/10.21037/atm.2016.03.36

Zhang, L., & Zheng, Y. (2018). Feedback as an assessment for learning tool: How useful can it be? *Assessment & Evaluation in Higher Education, 43*(7), 1120-1132. https://doi.org/10.1080/02602938.2018.1434481

Zhao, X., Liu, J. S., & Deng, K. (2013). Assumptions behind inter-coder reliability indices. In C. T. Salmon (Ed.), *Communication yearbook* (pp. 419-480). Taylor & Francis.

Zolfaghari, F., & Ahmadi, A. (2016). Assessment literacy components across subject matters. *Cogent Education, 3*(1), 1-16. http://doi.org/10.1080/2331186X.2016.1252561

Appendix A: List of Articles Included in the Literature Review

Aksit, O., & Wiebe, E. N. (2019). Exploring force and motion concepts in middle grades using computational modeling: A classroom intervention study. *Journal of Science Education and Technology*, 1-18.

Aliberas, J., Gutiérrez, R., & Izquierdo, M. (2019). Identifying changes in a student's mental models and stimulating intrinsic motivation for learning during a dialogue regulated by the teachback technique: A case study. *Research in Science Education*, 1-29.

Baek, H., & Schwarz, C. V. (2015). The influence of curriculum, instruction, technology, and social interactions on two fifth-grade students' epistemologies in modeling throughout a model-based curriculum unit. *Journal of Science Education and Technology, 24*(2-3), 216-233.

Bamberger, Y. M., & Davis, E. A. (2013). Middle-school science students' scientific modelling performances across content areas and within a learning progression. *International Journal of Science Education, 35*(2), 213-238.

Baumfalk, B., Bhattacharya, D., Vo, T., Forbes, C., Zangori, L., & Schwarz, C. (2019). Impact of model-based science curriculum and instruction on elementary students' explanations for the hydrosphere. *Journal of Research in Science Teaching*, *56*(5), 570-597.

Becker, S., & Jacobsen, M. (2019). "How can I build a model if I don't know the answer to the question?": Developing student and teacher sky scientist ontologies through making. *International Journal of Science and Mathematics Education*, *17*(1), 31-48.

Bielik, T., Fonio, E., Feinerman, O., Duncan, R. G., & Levy, S. T. (2020). Working together: Integrating computational modeling approaches to investigate complex phenomena. *Journal of Science Education and Technology*, 1-18.

Bouwma-Gearhart, J., Stewart, J., & Brown, K. (2009). Student misapplication of a gas-like model to explain particle movement in heated solids: Implications for curriculum and instruction towards students' creation and revision of accurate explanatory models. *International Journal of Science Education*, *31*(9), 1157-1174.

Brady, C., Holbert, N., Soylu, F., Novak, M., & Wilensky, U. (2015). Sandboxes for model-based inquiry. *Journal of Science Education and Technology, 24*(2-3), 265-286.

Bridle, C. A., & Yezierski, E. J. (2011). Evidence for the effectiveness of inquiry-based, particulate-level instruction on conceptions of the particulate nature of matter. *Journal of Chemical Education*, *89*(2), 192-198.

Buckley, B. C., Gobert, J. D., Horwitz, P., & O'Dwyer, L. (2010). Looking inside the black box: assessing model-based learning and inquiry in BioLogica™ . *International Journal of Learning Technology, 5*(2), 166-190.

Campbell, T., Oh, P. S., & Neilson, D. (2012). Discursive modes and their pedagogical functions in model-based inquiry (MBI) classrooms. *International Journal of Science Education, 34*(15), 2393-2419.

Campbell, T., Gray, R., & Fazio, X. (2019). Representing scientific activity: Affordances and constraints of central design and enactment features of a model-based inquiry unit. *School Science and Mathematics*, *119*(8), 475-486.

Campbell, T., McKenna, T. J., An, J., & Rodriguez, L. (2019). A responsive methodological construct for supporting learners' developing modeling competence in modeling-based learning environments. In *Towards a Competence-Based View on Models and Modeling in Science Education* (pp. 201-218). Springer, Cham.

Carpenter, S. L., Iveland, A., Moon, S., Hansen, A. K., Harlow, D. B., & Bianchini, J. A. (2019). Models are a "metaphor in your brain": How potential and preservice teachers understand the science and engineering practice of modeling. *School Science and Mathematics*, *119*(5), 275-286.

Chang, H. Y., & Chang, H. C. (2013). Scaffolding students' online critiquing of expert-and peer-generated molecular models of chemical reactions. *International Journal of Science Education, 35*(12), 2028-2056.

Cheng, M. F., & Brown, D. E. (2015). The role of scientific modeling criteria in advancing students' explanatory ideas of magnetism. *Journal of Research in Science Teaching, 52*(8), 1053-1081.

Cheng, M. F., & Lin, J. L. (2015). Investigating the relationship between students' views of scientific models and their development of models. *International Journal of Science Education*, *37*(15), 2453-2475.

Cisterna, D., Forbes, C. T., & Roy, R. (2019). Model-based teaching and learning about inheritance in third-grade science. *International Journal of Science Education*, *41*(15), 2177-2199.

Coll, R. K., & Treagust, D. F. (2001). Learners' mental models of chemical bonding. *Research in Science Education*, *31*(3), 357-382.

Cuperman, D., & Verner, I. M. (2019). Fostering analogical reasoning through creating robotic models of biological systems. *Journal of Science Education and Technology, 28*(2), 90-103.

Dauer, J. T., Bergan-Roller, H. E., King, G. P., Kjose, M., Galt, N. J., & Helikar, T. (2019). Changes in students' mental models from computational modeling of gene regulatory networks. *International Journal of STEM Education*, *6*(1), 38.

Demir, A., & Namdar, B. (2019). The effect of modeling activities on grade 5 students' informal reasoning about a real-life issue. *Research in Science Education*, 1-14.

Demirhan, E., & Şahin, F. (2019). The effects of different kinds of hands-on modeling activities on the academic achievement, problem-solving skills, and scientific creativity of prospective science teachers. *Research in Science Education*, 1-19.

Dickes, A. C., Sengupta, P., Farris, A. V., & Basu, S. (2016). Development of mechanistic reasoning and multilevel explanations of ecology in third grade using agent-based models. *Science Education*, *100*(4), 734-776.

Dolphin, G., & Benoit, W. (2016). Students' mental model development during historically contextualized inquiry: How the 'tectonic plate'metaphor impeded the process. *International Journal of Science Education, 38*(2), 276-297.

Duncan, R. G., Castro-Faix, M., & Choi, J. (2016). Informing a learning progression in genetics: Which should be taught first, Mendelian inheritance or the central dogma of molecular biology? *International Journal of Science and Mathematics Education, 14*(3), 445-472.

Evagorou, M., Nicolaou, C., & Lymbouridou, C. (2020). Modelling and argumentation with elementary school students. *Canadian Journal of Science, Mathematics and Technology Education*, 1-16.

Forbes, C. T., Lange-Schubert, K., Böschl, F., & Vo, T. (2019). Supporting primary students' developing modeling competence for water systems. In *Towards a Competence-Based View on Models and Modeling in Science Education* (pp. 257-273).

Fortus, D., Shwartz, Y., & Rosenfeld, S. (2016). High school students' meta-modeling knowledge. *Research in Science Education*, *46*(6), 787-810.

Fretz, E. B., Wu, H. K., Zhang, B., Davis, E. A., Krajcik, J. S., & Soloway, E. (2002). An investigation of software scaffolds supporting modeling practices. *Research in Science Education*, *32*(4), 567-589.

Galperin, D., & Raviolo, A. (2019). Reference frames and astronomy teaching: the development of a topocentric approach to the lunar phases. *Science Education International, 30*(1), 28.37.

Gobert, J. D. (2000). A typology of causal models for plate tectonics: Inferential power and barriers to understanding. *International Journal of Science Education, 22*(9), 937-977.

Gray, R., & Rogan-Klyve, A. (2018). Talking modelling: examining secondary science teachers' modelling-related talk during a model-based inquiry unit. *International Journal of Science Education, 40*(11), 1345-1366.

Gülen, S. (2020). A study to determine the ability of fifth-grade students in reflecting their knowledge about sun, earth, and moon by different measurement tools. *Science Education International*, *31*(1), 41-51.

Günther, S. L., Fleige, J., zu Belzen, A. U., & Krüger, D. (2019). Using the case method to foster preservice biology teachers' content knowledge and pedagogical content knowledge related to models and modeling. *Journal of Science Teacher Education, 30*(4), 321-343.

Guy-Gaytán, C., Gouvea, J. S., Griesemer, C., & Passmore, C. (2019). Tensions between learning models and engaging in modeling. *Science & Education*, 1-22.

Han, M., & Kim, H. B. (2019). Elementary students' modeling using analogy models to reveal the hidden mechanism of the human respiratory system. *International Journal of Science and Mathematics Education, 17*(5), 923-942.

Harlow, D. B., Bianchini, J. A., Swanson, L. H., & Dwyer, H. A. (2013). Potential teachers' appropriate and inappropriate application of pedagogical resources in a model-based physics course: A "knowledge in pieces" perspective on teacher learning. J*ournal of Research in Science Teaching, 50(*9), 1098-1126.

Heijnes, D., van Joolingen, W., & Leenaars, F. (2018). Stimulating scientific reasoning with drawing-based modeling. *Journal of Science Education and Technology, 27*(1), 45-56.

Hernández, M. I., Couso, D., & Pintó, R. (2015). Analyzing students' learning progressions throughout a teaching sequence on acoustic properties of materials with a model-based inquiry approach. *Journal of Science Education and Technology*, *24*(2-3), 356-377.

Hester, S. D., Nadler, M., Katcher, J., Elfring, L. K., Dykstra, E., Rezende, L. F., & Bolger, M. S. (2018). Authentic inquiry through modeling in biology (AIM-Bio): An introductory laboratory curriculum that increases undergraduates' scientific agency and skills. *CBE—Life Sciences Education, 17*(4), 1-23.

Hokayem, H., & Schwarz, C. (2014). Engaging fifth graders in scientific modeling to learn about evaporation and condensation. *International Journal of Science and Mathematics Education, 12*(1), 49-72.

Hsu, Y. S., Lai, T. L., & Hsu, W. H. (2015). A design model of distributed scaffolding for inquiry-based learning. *Research in Science Education, 45*(2), 241-273.

Jenkins, J. L., & Howard, E. M. (2019). Implementation of Modeling Instruction in a high school chemistry unit on energy and states of matter. *Science Education International, 30*(2), 97-104.

Jimenez-Liso, M. R., Martinez-Chico, M., Avraamidou, L., & López-Gay Lucio-Villegas, R. (2019). Scientific practices in teacher education: the interplay of sense, sensors, and emotions. *Research in Science & Technological Education*, 1-24.

Jong, J. P., Chiu, M. H., & Chung, S. L. (2015). The use of modeling-based text to improve students' modeling competencies. *Science Education, 99*(5), 986-1018.

Justi, R., Gilbert, J. K., & Ferreira, P. F. (2009). The application of a 'model of modelling' to illustrate the importance of metavisualisation in respect of the three types of representation. In *Multiple representations in chemical education* (pp. 285-307). Springer.

Kawasaki, J., & Sandoval, W. A. (2020). Examining teachers' classroom strategies to understand their goals for student learning around the science practices in the Next Generation Science Standards. *Journal of Science Teacher Education*, *31*(4), 1-17.

Ke, L., & Schwarz, C. V. (2019). Using epistemic considerations in teaching: Fostering students' meaningful engagement in scientific modeling. In *Towards a Competence-Based View on Models and Modeling in Science Education* (pp. 181-199). Springer.

Kenyon, L., Davis, E. A., & Hug, B. (2011). Design approaches to support preservice teachers in scientific modeling. *Journal of Science Teacher Education, 22*(1), 1-21.

Khan, S. (2007). Model-based inquiries in chemistry. *Science Education, 91*(6), 877-905.

Khan, S. (2008a). Co-construction and model evolution in chemistry. In J. J. Clement, M. A. Rea-Ramirez (Eds.), *Model based learning and instruction in science* (pp. 59-78). Springer.

Khan, S. (2008b). What if scenarios for testing student models in chemistry. In J. J. Clement, M. A. Rea-Ramirez (Eds.), *Model based learning and instruction in science* (pp. 139-150). Springer.

Khan, S. (2011a). What's missing in model-based teaching. *Journal of Science Teacher Education, 22*(6), 535-560.

Khan, S. (2011b). New pedagogies on teaching science with computer simulations. *Journal of Science Education and Technology, 20*(3), 215-232.

King, G. P., Bergan-Roller, H., Galt, N., Helikar, T., & Dauer, J. T. (2019). Modelling activities integrating construction and simulation supported explanatory and evaluative reasoning. *International Journal of Science Education*, *41*(13), 1-23.

Lally, D., & Forbes, C. (2019). Modelling water systems in an introductory undergraduate course: Students' use and evaluation of data-driven, computer-based models. *International Journal of Science Education, 41*(14), 1999-2023.

Lamar, M. F., Wilhelm, J. A., & Cole, M. (2018). A mixed methods comparison of teachers' lunar modeling lesson implementation and student learning outcomes. *The Journal of Educational Research, 111*(1), 108-123.

Lee, S., & Kim, H. B. (2014). Exploring secondary students' epistemological features depending on the evaluation levels of the group model on blood circulation. *Science & Education, 23*(5), 1075-1099.

Lee, Y. C. (2015). Self-generated analogical models of respiratory pathways. *Journal of Biological Education, 49*(4), 370-384.

Louca, L. T., & Zacharias, Z. C. (2015). Examining learning through modeling in K-6 science education. *Journal of Science Education and Technology, 24*(2-3), 192-215.

Louca, L. T., Zacharias, Z. C., & Constantinou, C. P. (2011). In Quest of productive modeling-based learning discourse in elementary school science. *Journal of Research in Science Teaching, 48*(8), 919-951.

Maia, P. F., & Justi, R. (2009). Learning of chemical equilibrium through modelling-based teaching. *International Journal of Science Education, 31*(5), 603-630.

Mendonça, P. C. C., & Justi, R. (2011). Contributions of the model of modelling diagram to the learning of ionic bonding: Analysis of a case study. *Research in Science Education*, *41*(4), 479-503.

Mendonça, P. C. C., & Justi, R. (2013). The relationships between modelling and argumentation from the perspective of the model of modelling diagram. *International Journal of Science Education, 35*(14), 2407-2434.

Mendonça, P. C. C., & Justi, R. (2014). An instrument for analyzing arguments produced in modeling-based chemistry lessons. *Journal of Research in Science Teaching, 51*(2), 192-218.

Merritt, J., & Krajcik, J. (2013). Learning progression developed to support students in building a particle model of matter. In *Concepts of matter in science education* (pp. 11-45). Springer.

Mierdel, J., & Bogner, F. X. (2019). Investigations of modellers and model Viewers in an out-of-school gene technology laboratory. *Research in Science Education*, 1-22.

Nelson, M. M., & Davis, E. A. (2012). Preservice Elementary Teachers' Evaluations of Elementary Students' Scientific Models: An aspect of pedagogical content knowledge for scientific modeling. *International Journal of Science Education, 34*(12), 1931-1959.

Nielsen, S. S., & Nielsen, J. A. (2019). A competence-oriented approach to models and modelling in lower secondary science education: Practices and rationales among danish teachers. *Research in Science Education*, 1-29.

Nunez-Oviedo, M. C., & Clement, J. J. (2019). Large Scale Scientific Modeling Practices That Can Organize Science Instruction at the Unit and Lesson Levels. *Development of Student Understanding: Focus on Science Education, 4*(68), 1-22.

Nunez-Oviedo, M. C., & Clement, J. (2008). A competition strategy and other modes for developing mental models in large group discussion. In *Model based learning and instruction in science* (pp. 117-138). Springer.

Oh, P. S. (2010). How can teachers help students formulate scientific hypotheses? Some strategies found in abductive inquiry activities of Earth Science. *International Journal of Science Education*, *32*(4), 541-560.

Oh, P. S. (2019). Features of modeling-based abductive reasoning as a disciplinary practice of inquiry in earth science. *Science & Education*, *28*(6-7), 731-757.

Passmore, C., & Stewart, J. (2002). A modeling approach to teaching evolutionary biology in high schools. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, *39*(3), 185-204.

Passmore, C. M., & Svoboda, J. (2012). Exploring opportunities for argumentation in modelling classrooms. *International Journal of Science Education, 34*(10), 1535-1554.

Peel, A., Zangori, L., Friedrichsen, P., Hayes, E., & Sadler, T. (2019). Students' model-based explanations about natural selection and antibiotic resistance through socio-scientific issues-based learning. *International Journal of Science Education, 41*(4), 510-532.

Pierson, A. E., & Clark, D. B. (2019). Sedimentation of Modeling Practices. *Science & Education*, *28*(8), 1-29.

Pierson, A. E., Brady, C. E., & Clark, D. B. (2020). Balancing the environment: Computational models as interactive participants in a STEM classroom. *Journal of Science Education and Technology*, *29*(1), 101-119.

Pluta, W. J., Chinn, C. A., & Duncan, R. G. (2011). Learners' epistemic criteria for good scientific models. *Journal of Research in Science Teaching*, *48*(5), 486-511.

Prins, G. T., Bulte, A. M., & Pilot, A. (2011). Evaluation of a design principle for fostering students' epistemological views on models and modelling using authentic practices as contexts for learning in chemistry education. *International Journal of Science Education*, *33*(11), 1539-1569.

Raghavan, K., Sartoris, M. L., & Glaser, R. (1998). Why does it go up? The impact of the MARS curriculum as revealed through changes in student explanations of a helium balloon. *Journal of Research in Science úınTeaching, 35*(5), 547-567.

Rea-Ramirez, M. A., Clement, J., & Nunez-Oviedo, M. C. (2008). Model based reasoning among inner city middle school students. In *Model based learning and instruction in science* (pp. 233-253). Springer.

Reinagel, A., & Bray Speth, E. (2016). Beyond the central dogma: model-based learning of how genes determine phenotypes. *CBE—Life Sciences Education, 15*(1), 1-13.

Ruppert, J., Duncan, R. G., & Chinn, C. A. (2019). Disentangling the role of domain-specific knowledge in student modeling. *Research in Science Education, 49*(3), 921-948.

Ryu, S., Han, Y., & Paik, S. H. (2015). Understanding co-development of conceptual and epistemic understanding through modeling practices with mobile internet. *Journal of Science Education and Technology, 24*(2-3), 330-355.

Samarapungavan, A., Bryan, L., & Wills, J. (2017). Second graders' emerging particle models of matter in the context of learning through model-based inquiry. *Journal of Research in Science Teaching, 54*(8), 988-1023.

Schwarz, C. (2009). Developing preservice elementary teachers' knowledge and practices through modeling-centered scientific inquiry. *Science Education, 93*(4), 720-744.

Schwarz, C. V., & Gwekwerere, Y. N. (2007). Using a guided inquiry and modeling instructional framework (EIMA) to support preservice K-8 science teaching. *Science education*, *91*(1), 158-186.

Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Achér, A., Fortus, D., Shwartz, Y., & Krajcik, J. (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *Journal of Research in Science Teaching, 46*(6), 632-654.

Schwarz, C., Reiser, B. J., Acher, A., Kenyon, L., & Fortus, D. (2012). MoDeLS: Challenges in defining a learning progression for scientific modeling. In *Learning progressions in science* (pp. 101-137). Brill Sense.

Shemwell, J. T., & Capps, D. K. (2019). Learning abstraction as a modeling competence. In *Towards a Competence-Based View on Models and Modeling in Science Education* (pp. 291-307). Springer.

Sherwood, C. A. (2020). "The goals remain elusive": Using drawings to examine shifts in teachers' mental models before and after an NGSS Professional learning experience. *Journal of Science Teacher Education*, 1-23.

Spier-Dance, L., Mayer-Smith, J., Dance, N., & Khan, S. (2005). The role of student-generated analogies in promoting conceptual understanding for undergraduate chemistry students. *Research in Science & Technological Education*, *23*(2), 163-178.

Sung, J. Y., & Oh, P. S. (2018). Sixth grade students' content-specific competencies and challenges in learning the seasons through modeling. *Research in Science Education, 48*(4), 839-864.

Svoboda, J., & Passmore, C. (2013). The strategies of modeling in biology education. *Science & Education, 22*(1), 119-142.

Tay, S. L., & Yeo, J. (2018). Analysis of a physics teacher's pedagogical 'micro-actions' that support 17-year-olds' learning of free body diagrams via a modelling approach. *International Journal of Science Education*, *40*(2), 109-138.

Thompson, J. J., Hagenah, S., McDonald, S., & Barchenger, C. (2019). Toward a practice-based theory for how professional learning communities engage in the improvement of tools and practices for scientific modeling. *Science Education, 103*(6), 1423-1455.

van Joolingen, W. R., Schouten, J., & Leenaars, F. (2019). Drawing-based modeling in teaching elementary biology as a diagnostic tool. In *Towards a Competence-Based View on Models and Modeling in Science Education* (pp. 131-145). Springer, Cham.

Vasconcelos, L., & Kim, C. (2019). Coding in scientific modeling lessons (CS-ModeL). *Educational Technology Research and Development*, 1-27.

Vergara-Díaz, C., Bustamante, K., Pinto, L., & Cofré, H. (2020). Exploring Chilean seventh grade students' conceptions of Earth dynamics before and after model-and inquiry-based instruction. *Journal of Geoscience Education*, 1-11.

Vo, T., Forbes, C., Zangori, L., & Schwarz, C. V. (2019). Longitudinal investigation of primary inservice teachers' modelling the hydrological phenomena. *International Journal of Science Education*, *41*(18), 1-20.

Werner, S., Förtsch, C., Boone, W., von Kotzebue, L., & Neuhaus, B. J. (2019). Investigating how German biology teachers use three-dimensional physical models in classroom instruction: A video study. *Research in Science Education, 49(*2), 437-463.

Wilensky, U., & Reisman, K. (2006). Thinking like a wolf, a sheep, or a firefly: Learning biology through constructing and testing computational theories—an embodied modeling approach. *Cognition and instruction*, *24*(2), 171-209.

Wilkerson-Jerde, M. H., Gravel, B. E., & Macrander, C. A. (2015). Exploring shifts in middle school learners' modeling activity while generating drawings, animations, and computational simulations of molecular diffusion. *Journal of Science Education and Technology, 24*(2-3), 396-415.

Wilkerson, M. H., Shareff, R., Laina, V., & Gravel, B. (2018). Epistemic gameplay and discovery in computational model-based inquiry activities. *Instructional Science, 46*(1), 35-60.

Williams, G., & Clement, J. (2015). Identifying multiple levels of discussion-based teaching strategies for constructing scientific models. *International Journal of Science Education, 37*(1), 82-107.

Windschitl, M., Thompson, J., & Braaten, M. (2008). How novice science teachers appropriate epistemic discourses around model-based inquiry for use in classrooms. *Cognition and Instruction*, *26*(3), 310-378.

Xiang, L., & Passmore, C. (2015). A framework for model-based inquiry through agent-based programming. *Journal of Science Education and Technology, 24*(2-3), 311-329.

Yeşiloğlu, S. N. (2019). Investigation of pre-service chemistry teachers' understanding of radioactive decay: a laboratory modelling activity. *Chemistry Education Research and Practice*, *20*(4), 862-872.

Zangori, L., & Forbes, C. T. (2015). Exploring third-grade student model-based explanations about plant relationships within an ecosystem. *International Journal of Science Education*, *37*(18), 2942-2964.

Zangori, L., & Forbes, C. T. (2016). Development of an empirically based learning performances framework for third-grade students' model-based explanations about plant processes. *Science Education, 100*(6), 961-982.

Zangori, L., Forbes, C. T., & Schwarz, C. V. (2015). Exploring the effect of embedded scaffolding within curricular tasks on third-grade students' model-based explanations about hydrologic cycling. *Science & Education, 24*(7-8), 957-981.

Zangori, L., Vo, T., Forbes, C. T., & Schwarz, C. V. (2017). Supporting 3rd-grade students model-based explanations about groundwater: a quasi-experimental study of a curricular intervention. *International Journal of Science Education, 39*(11), 1421-1442.

Zwickl, B. M., Hu, D., Finkelstein, N., & Lewandowski, H. J. (2015). Model-based reasoning in the physics laboratory: Framework and initial results. *Physical Review Special Topics-Physics Education Research*, *11*(2), 020113.

Appendix B: Full Version of the QALMBT

Invitation and consent form

Project: Exploring in-service science teachers' assessment literacy

Principal Investigator:
Dr. Samia Khan (UBC), Associate Professor, Department of Curriculum and Pedagogy, University of British Columbia, 2125 Main Mall, UBC, Vancouver, BC, Canada V6T 1Z4

Co-Investigator:
Alexis Gonzalez, PhD. candidate

You are being invited to participate in a study called "Exploring in-service science teachers' assessment literacy". The purpose of this study is to investigate in-service science teachers' (ISTs) understanding of how to teach and assess students. The goal is to identify existing teaching strategies that ISTs implement in their pedagogy to help students reflect on the nature of scientific knowledge and improve their understanding of core ideas in science.

You are being invited to take part in this research study because you are an in-service science teacher who is currently teaching in science classrooms. Please read carefully the attached consent form.
Once you agree to participate in this research by selecting the checkbox at the bottom of this consent form, you will immediately have access to a 20/30-minute close-ended study questionnaire.

If you have any questions or concerns about this study, please contact the co-investigator. Alexis Gonzalez, Ph.D. student at the University of British Columbia
If you have any concerns in respect of your rights as a research participant, you can contact the the Research Participant Complaint Line in the UBC Office of Research Ethics at 604-822-8598 or if long distance e-mail RSIL@ors.ubc.ca or call toll free 1-877-822-8598.

Sincerely,
Alexis Gonzalez
Ph.D. Student in Curriculum Studies
University of British Columbia

****

Consent signature

Taking part in this study is totally voluntary. You have the right to approve the use of your results for the research.

Your signature below confirms that you have been informed about the study and that you consent to participate.

1. I consent to be part in this study and I am aware of the details of the research.
2. I have been informed that the purpose of this research is to investigate in-service science teachers' understanding about how to teach and assess students in the science classroom.
3. I understand that my participation in this research is for research purposes.
4. I have been informed of the possible risks and benefits of participating in this research.
5. In this study I will be asked to answer a questionnaire.
6. I acknowledge that only if I consent to participate in the second phase of the study I may be contacted to be part of the following stages: (i) interview (audio-recorded), (ii) class observations (at least 5 classes which will be video recorded) before attending an online professional development course in science education, (iii) taking an online course in science education (approximately 10 hours), (iv) class observations of a unit (at least 5 classes which will be video recorded), and (v) a post interview (audio recorded).
7. I know that I can withdraw from this research at any time and my decision will be respected.
8. I understand that the data from this study will be stored digitally by the principal investigator and confidentiality will be safeguarded by altering personal information and using passwords. I also understand that the data will be encrypted.
9. I have been informed that I may request to review my individual results (e.g., interview transcripts).
10. I understand that by signing below, I allow the researchers to collect data.

☐ Please select this box to consent for the first phase of the study.

☐ Please select this box if you agree to have the possibility to participate in the second phase

Date:
DD/MM/YYYY

0% 100%

Project: Exploring in-service science teachers' assessment literacy

Dear science teacher:

Thank you for your interest in participating in this study. The following questionnaire will identify how you teach science and how you assess students. There are no right or wrong answers, just descriptions and characteristics. The questionnaire includes three parts. The first part is a brief section which will be used to characterize participants, identify your disciplinary background and level of education. The second part is longer and asks questions regarding how you teach science in terms of your pedagogy. Finally, the last section explores your understanding of the nature and purpose of models and modeling.

Please respond to every question. Also, in the last part of the questionnaire, you will have the option to comment (if you desire) on any of the question items.

Thank you very much for your participation.

→

I. Demographic information

Indicate the province in which you teach:

1. Please provide your name: (* we will only use your personal name to contact you in case you agree to participate in the second phase of this study).

2. Name of the school in which you teach:

3. City of the school:

4. Please provide your email: (You might be contacted for the next phase only if you provided consent in the consent form. Moreover, this information will be used to contact you in case you are a winner of the book in science education).

5. Personal Identifier: (Please create a personal identifier which will be used instead of your real name).

The last two letters of your first name	
The first two letters of your school's name	
The two first letters where your school is located	
The two last numbers of the year in which you obtained your Bachelor of Education	

6. Gender.

7. What is your disciplinary background? If you have more than one primary field, please select each of them.

- ☐ Physics
- ☐ Chemistry
- ☐ Biology
- ☐ Geology
- ☐ Engineering
- ☐ Other (please specify your disciplinary background if it is not included in the options)

8. What is your highest level of education (finished program)? If you have the same level of degree more than once, e.g., MSc and MEd, select both degrees in the checkbox list.

- ☐ Bachelor's Degree
- ☐ Master of Education
- ☐ PhD in Science Education
- ☐ Master of Science
- ☐ PhD in Science
- ☐ Other (please specify your highest degree if it is not included)

9. Indicate your bachelor degree. If you have more than one bachelor degree, please select the options that better represent each degree.

- ☐ Administration Studies
- ☐ Architecture Studies
- ☐ Art Studies
- ☐ Business Studies
- ☐ Economic Studies
- ☐ Education
- ☐ Engineering Studies
- ☐ Environmental Studies
- ☐ Journalism and Mass Communication
- ☐ Law Studies
- ☐ Life Sciences
- ☐ Management Studies
- ☐ Marketing Studies
- ☐ Political Sciences
- ☐ Psychology
- ☐ Sociology
- ☐ Sustainability Studies
- ☐ Other

10. Name of the university where you finished your teacher education program. If your university is not listed, select "other" and specify the name of the institution.

- O Simon Fraser University
- O Trinity Western University
- O Thompson Rivers University
- O University of the Fraser Valley
- O University of British Columbia
- O University of British Columbia-Okanagan
- O University of Northern British Columbia
- O University of Vicotria
- O Vancouver Island University
- O University of Calgary
- O University of Toronto
- O University of McGill
- O Unversity of Montreal
- O University of Alberta
- O University of Ottawa
- O Other (Please specify the name)

11. Select from the checkbox list whether your qualification is for middle school, secondary school, or both.

__ Middle school (grades 6-8)

__ Secondary school (grades 9-12)

__ Both (Middle and Secondary school)

12. Indicate the number of courses taken on **assessment** while studying your teacher education program. (If you attended courses that covered the topic of assessment only as a portion of a course, e.g., a science method course, please indicate that information in the next question).

1, 2, 3, 4, 5, 6…

13. Indicate the number of courses while studying your teaching program in which you studied the topic of assessment only as a portion of the course and not as the main topic of the course.

1, 2, 3, 4, 5, 6…

14. From the following list, select if you have learned about the following topics in science courses, science teacher education courses, or professional development courses.

- ☐ Nature of science
- ☐ Models and modeling in science education
- ☐ Assessment in science education
- ☐ Assessing scientific reasoning
- ☐ Learning progression in science education
- ☐ Support teaching for conceptual understanding
- ☐ Strategies to elicit students' ideas
- ☐ The role of critical thinking and argumentation in science education
- ☐ None of them

17. Select the number of teaching hours a week (on average) that you have been teaching science over the last 3 years. If you just started teaching this year, indicate your current number of teaching hours.

18. On average, how many students are typically in one of your classes?

__ 1 to 5 __ 6 to 10 __ 11 to 15 __16 to 20 __ 21 to 25 __ 26 to 30 __ 31 to 35 __ 36 to 40

__ 41 or more

0% 100%

II. Your Pedagogy
(Please read carefully the instructions)
For each statement below mark the option which best reflects your perspective on your practice regarding how often you include it in your pedagogy:

1. Never
2. Very Rarely (only a couple of times per year at a maximum)
3. Sometimes (about once every month or once every unit)
4. Frequently (about once every two weeks in a unit)
5. Very frequently (almost every class

Please notice that each item is comprised of a common statement followed by two sentences:

- **The first sentence** focuses on your practice when you teach science **without emphasizing a particular teaching approach**

- **The second sentence** focuses on your practice specifically when you teach science with or about **scientific models in your pedagogy**.

We make this distinction because as science teachers, we use and teach with scientific models when teaching core ideas in the science classroom, such as ecosystems, climate change, energy transfer, equilibrium, atoms, and solar system.

Questionnaire:

1. When I assess students learning, I evaluate whether students understand that...

	1	2	3	4	5
... knowledge may change in light of new evidence.	O	O	O	O	O
... models can be refined based on new evidence.	O	O	O	O	O

2. I align my assessment with the goals of the provincial science curriculum when assessing...

	1	2	3	4	5
... students' ideas in science.	O	O	O	O	O
... the expected models that students should learn.	O	O	O	O	O

3. For those areas that students have difficulty in comprehending, I promote the generation of a consensus ...

	1	2	3	4	5
... explanation that helps students have a similar understanding of the core ideas.	O	O	O	O	O
... model that helps students have a similar understanding of the phenomenon under study.	O	O	O	O	O

4. I challenge my students to develop assessment criteria to evaluate...

	1	2	3	4	5
... their classmates' explanations or answers.	O	O	O	O	O
... the models constructed by their classmates.	O	O	O	O	O

5. I explain to students the criteria that I will use to evaluate...

	1	2	3	4	5
... their understanding.	O	O	O	O	O
... their models.	O	O	O	O	O

6. I use assessments to measure how students...

	1	2	3	4	5
... carry out investigations.	O	O	O	O	O
... develop models to guide their investigations.	O	O	O	O	O

7. I use results from an assessment to compare how students'...

	1	2	3	4	5
... understanding about the topic under study has been reshaped.	O	O	O	O	O
... ideas about a model have been reshaped.	O	O	O	O	O

8. When I develop summative assessment, I inform students in advance about the criteria that I will use to assess...

	1	2	3	4	5
... understanding in class.	O	O	O	O	O
... their models.	O	O	O	O	O

9. I design scaffolded assignments or tasks that progress in complexity in order to assess students' understanding about...

	1	2	3	4	5
... disciplinary core ideas in science.	O	O	O	O	O
... the model under study.	O	O	O	O	O

10. I develop different kinds of assessment to evaluate...

	1	2	3	4	5
... students' reasoning.	O	O	O	O	O
... students' reasoning by using models.	O	O	O	O	O

11. I ask students to test their hypothesis by...

	1	2	3	4	5
... identifying relationships between variables.	O	O	O	O	O
... using their model.	O	O	O	O	O

12. When I make an attempt for students to understand a...

	1	2	3	4	5
... core idea in science, I organize the content in my lessons following a sequence which considers how student understanding can evolve over a span of time.	O	O	O	O	O
... model, I organize the content in my lessons following a sequence which considers how student understanding can evolve over a span of time.	O	O	O	O	O

13. In my classes I ask students to comment on ...

	1	2	3	4	5
... their classmates' ideas or answers.	O	O	O	O	O
... the models created by their classmates.	O	O	O	O	O

14. When I assess students, I...

	1	2	3	4	5
... consider different levels of complexity to allow students progress in their understanding about a core idea.	O	O	O	O	O
... allow them to refine their models in order to help them reach different levels of complexity about the phenomenon that they are modeling.	O	O	O	O	O

15. I am able to translate the curriculum goals into clear specific tasks to

	1	2	3	4	5
... guide my assessment activities.	O	O	O	O	O
... evaluate students' models.	O	O	O	O	O

16. I communicate the results of the assessment in order to help each student

	1	2	3	4	5
... refine their initial ideas.	O	O	O	O	O
... achieve a better understanding of the expected model that I want them to learn.	O	O	O	O	O

17. I teach students how to judge the quality of their explanations based on the consistency of their...

	1	2	3	4	5
... ideas.	O	O	O	O	O
... models.	O	O	O	O	O

18. I deconstruct a task or objective from the science curriculum into smaller instructional learning experiences that assess students' progression with...

	1	2	3	4	5
... a core idea.	O	O	O	O	O
... their models.	O	O	O	O	O

0% 100%

Scale:
1. Never
2. Very Rarely (only a couple of times per year at a maximum)
3. Sometimes (about once every month or once every unit)
4. Frequently (about once every two weeks in a unit)
5. Very frequently (almost every class

19. After a summative assessment, I clarify...

	1	2	3	4	5
... students' wrong answers.	O	O	O	O	O
... common students' misconceptions or alternative ideas about their generated models in class.	O	O	O	O	O

20. In order to assess students, I include laboratory activities that...

	1	2	3	4	5
... reinforce students' understanding of the core ideas addressed in class.	O	O	O	O	O
... require the construction of models by students.	O	O	O	O	O

21. I use assessment to judge students' understanding about the phenomenon...

	1	2	3	4	5
... under study.	O	O	O	O	O
... to be modeled.	O	O	O	O	O

22. When students express their claims in front of the classroom, I establish classroom norms to promote a safe expression of students'...

	1	2	3	4	5
... ideas about the disciplinary core ideas in science.	O	O	O	O	O
... ideas about their models.	O	O	O	O	O

23. I design different scoring tools (e.g., rubrics, checklists, standards) to...

	1	2	3	4	5
... judge students work.	O	O	O	O	O
... evaluate the models generated by students.	O	O	O	O	O

24. I use the results generated from formative assessment to adjust the content of my lessons regarding...

	1	2	3	4	5
... the core ideas under study.	O	O	O	O	O
... the model that I expect student to learn.	O	O	O	O	O

25. I encourage students to use their pre-existing ideas in order to help them to construct an initial...

	1	2	3	4	5
... explanation that can be enriched later.	O	O	O	O	O
... model than can be enriched later.	O	O	O	O	O

26. I use assessments to measure students' understanding of...

	1	2	3	4	5
... core ideas.	O	O	O	O	O
... the models generated by them.	O	O	O	O	O

27. I assess how students make judgements in science based on...

	1	2	3	4	5
... reasoning.	O	O	O	O	O
... reasoning with a model.	O	O	O	O	O

28. When I develop distractors (incorrect or inferior alternatives) in a test, I have in mind the different alternatives or inaccurate ideas that students might have for the...

	1	2	3	4	5
... content that they are studying.	O	O	O	O	O
... model in question.	O	O	O	O	O

29. I challenge my students to show evidence...

	1	2	3	4	5
... to support their claims about the content that they are studying.	O	O	O	O	O
... to support their claims about their models.	O	O	O	O	O

30. I use the results of the assessment to coach a student when she/he/they are having problems understanding...

	1	2	3	4	5
... a disciplinary core idea in science.	O	O	O	O	O
... a model.	O	O	O	O	O

31. I use assessment to evaluate the internal consistency or coherence of...

	1	2	3	4	5
... students' ideas.	O	O	O	O	O
... various models constructed by a student.	O	O	O	O	O

32. When I develop assessment instruments to assess students'...

	1	2	3	4	5
... understanding, I think beforehand how I will interpret the results.	O	O	O	O	O
... models, I think beforehand how I will interpret the results	O	O	O	O	O

33. I use assessment to give formative feedback to students about...

	1	2	3	4	5
... their understanding of the core ideas studied in class.	O	O	O	O	O
... the phenomenon that they modeled.	O	O	O	O	O

34. I use assessment to locate evidence about the missing elements...

	1	2	3	4	5
... that students have not understood regarding the core ideas under study.	O	O	O	O	O
... in a model that students have not understood.	O	O	O	O	O

35. I tailor assessment in order to give all students the best opportunities to express their understanding about the...

	1	2	3	4	5
... disciplinary core ideas studied in class.	O	O	O	O	O
... model under study.	O	O	O	O	O

0% 100%

This is the last page of the questionnaire.

II. Models and Modeling

For each statement below mark to what extent you agree or disagree with the statement:
1. Strongly disagree
2. Disagree
3. Undecided
4. Agree
5. Strongly agree

	1	2	3	4	5
1. A model is a theoretical construction of reality.	O	O	O	O	O
2. Hypotheses can be tested by using a model.	O	O	O	O	O
3. Models need to be refined based on an iterative process in which empirical data compels the revision of the model.	O	O	O	O	O
4. Models can be designed with a main purpose of making testable predictions between variables.	O	O	O	O	O
5. Models could be changed when deduced hypotheses do not explain the original event, phenomena or object.	O	O	O	O	O
6. A model differs in some degree from the reality.	O	O	O	O	O
7. Testing competing scientific models gives better insight into the explanatory scope of each of them.	O	O	O	O	O
8. Models are used to help formulate ideas and theories about scientific events, phenomena or objects.	O	O	O	O	O
9. A model can be adjusted to reflect new findings.	O	O	O	O	O
10. A model can be tested conceptually or non-experimentally.	O	O	O	O	O
11. A characteristic of models is that they can be disproved when problems with its explanatory adequacy are identified by scientists.	O	O	O	O	O
12. Models are developed to allow us to raise new questions and create new problems.	O	O	O	O	O
13. A model can be a mental image about a phenomenon that represents some entities of the original object under study.	O	O	O	O	O
14. Competing models can coexist in science to represent the same object, phenomenon or system.	O	O	O	O	O
15. Models can serve to transfer findings about a specific idea to another phenomenon.	O	O	O	O	O
16. Models can be used to test assumptions about something.	O	O	O	O	O
17. Models need to be assessed to test their validity and fit with reality.	O	O	O	O	O
18. A model is a research tool that can be used to generate information.	O	O	O	O	O
19. A model can represent a specific part of a phenomenon under study rather than representing the entire phenomenon that constitutes in the real world.	O	O	O	O	O
20. Each model has limitations making it necessary to generate several models to represent the reality.	O	O	O	O	O

Appendix C: Example of Back translation of the QALMBT-Generic and -Modeling

Comparison between the English version of the QALMBT questionnaire and the Spanish version. A back translation of the instrument is provided in order to show the rewording of the items to ensure semantic and conceptual equivalence. (UA: unaltered; LA: little altered with same meaning, A: Altered with a similar meaning)

Item	Original	Back translation of the translation suggested by the author Ph.D. Biotechnology with experience teacher in undergraduate programs in Chile.	Back translation of the translation suggested by the external researcher Master in Education with experience teaching in Chile and the United States.	Spanish translation (Author)	Spanish translation (External researcher) Ph.D. Science Education with teaching experience in Mexico.	
1	1. When I assess students learning, I evaluate whether students understand that… ... knowledge may change in light of new evidence ... models can be refined based on new evidence.	When I evaluate the student learning, I evaluate if the students understand that... ... knowledge may change based on new evidence. ... the models can be refined in relation to new evidence.	When evaluating students' learning, I evaluate if students understand… …..that knowledge can change when new evidence might become available. … new models might sophisticate themselves when new evidence is discovered	Cuando evalúo el aprendizaje de los estudiantes, evalúo si los estudiantes entienden que... … el conocimiento puede cambiar en función de nueva evidencia. … los modelos pueden ser refinados en relación a una nueva evidencia.	Cuando evalúo el aprendizaje de los estudiantes, yo evalúo si los estudiantes entienden que… … el conocimiento puede cambiar a la luz de nuevas evidencias …los modelos pueden sofisticarse en base a nuevas evidencias.	UA LA A

Appendix D: Example of Items Revised by the External Researcher

Comments discussed with the external researcher during the process of revision of the translated items. Only comments for those items that were necessary to discuss are included in the table. The comments are related to the comparison between both Spanish versions translated for each researcher.

Item revised QALMBT-Generic/Modeling	Authors' comment	External researcher's comment
1	I think refined or sophisticated have a similar meaning in Spanish. I prefer to keep "refined" (refinado) to keep the translation literal.	I agree.
2	I explicitly included the Ministry of Education because in Chile is the ministry the entity that determines and suggests the science curriculum. This situation is similar to the British Columbia' science curriculum which is suggested by the province.	In Mexico we have a "Secretaría de Education" or Secretariat of Education. I think that for the survey's purposes you don't really need to mention the ministry because the focus of this question is the action: if the teachers align normative learning goals and their planning.
4	In Chile, "Retar" has a negative connotation. It means "challenge" but also "regañar" (to nag).	Okay.
7	I preferred to include two words "enriched and modified" to cover the meaning related to "reshaped". I think redefined in the Spanish context (2nd translation) might be a little bit confusing even though is a more literal translation.	I agree.
8	I included the words "con nota" (grade) because science teachers who only have a bachelor in science but not a degree in education might not know the word. I preferred to clearly indicate the constructions of models to enhance the coherence in the sentence.	Okay.

Appendix E: Full Interview Protocol

First Interview:

From an educational point of view, what were the main reasons that motivated you to participate in this study?

I: In general, what is your approach to planning your classes, for example, in terms of content selection, activities, and evaluation?

I: What strategies do you use in your class to motivate students to participate in it? I: How do you make sure that the students understood or did not understand the content that you wanted them to learn?

I: What is your main objective (purpose) when evaluating the students in your class?

I: What instructional strategies, e.g., tests, rubric, formative assessment, do you generally use to assess students in terms of content and skills?

E: The following questions are related to the use of models in your pedagogy. As science teachers, we use and teach with scientific models when we teach core ideas in the science classroom, such as ecosystems, climate change, energy transfer, chemical equilibrium, atoms, and the solar system. Before starting the first observation cycle, I asked you to choose a unit in which you have the possibility to include models. In this sense,

I: Do you include models in your teaching? Could you give me an example?

I: How do you motivate your students to create their own models? I: When students build a model, what is your objective of the activity?

I: What aspects and characteristics do you emphasize when evaluating modeling?

I: How do you assess student modeling at the beginning, during and at the end of a unit?

I: In terms of teaching planning and how you assess students, before planning and conducting your class, how do you translate science curriculum objectives into specific tasks to guide your assessment activities?

I: How do you use evaluation to assess whether the model built by the student is aligned with the model you want them to learn?

I: Once you have asked the students to think with models, what criteria do you use to evaluate the student models?
I: When you evaluate your students, how do you generally communicate formative feedback to students? I: Do you use the same strategies whether or not the activity involves modeling? How is it different?

I: When students work with models. What is your role in class and how do students interact with you and each other while using models?

I: After an evaluation, how do you generally use the results of the evaluation to adjust your pedagogy or how you teach?

I: If the performance of the students is not as expected, do you adapt or reduce the complexity of the model that you want the students to learn? Could you indicate what adaptations you make?

I: How do you involve or involve your students in the process of evaluating or judging the quality of the models or explanations of their classmates?

I: How do you use assessment to monitor a student when they have trouble understanding a model?

I: Could you describe how you give students opportunities to express their understanding of the model you want them to learn at the beginning, during and at the end of a unit?

I: How do you organize the content of your classes within a unit once you have already selected the model you want to teach?

I: Do you incorporate students' previous ideas throughout a unit when they study some model in science? How do you evaluate how these previous ideas change in the unit?

I: How do you assess the students' progress in their understanding of the model in a unit?

I: And finally, how do you use assessment to help students enrich and refine their inquiry skills when they think with models?

Second Interview:

I: To begin this interview, I would like to ask you, do you think your ideas about pedagogy could have changed in any way after completing the modules? How?

I: Do you think the model-based course was helpful in understanding how to improve students' conceptual understanding by creating models?

I: Did the structure of the modules make sense to you? What strengths and weaknesses did you find?

I: Did you use any strategies in your class to guide the students in the analysis of evidence?
I: Do you think MBT is an effective approach to teaching each of the contents of the science curriculum?

I: Based on your experience implementing MBT, what are the advantages and disadvantages of this approach?

Now, I would like to ask you questions about how you assess the students in your class using models and how your ideas might have changed after taking the OPDC. In the event that you cannot answer any question because you could not incorporate an element related to modeling, indicate that you did not do it and indicate how you would do it now that you already have a general knowledge about models and modeling in science)

I: In what way do you think your ideas about evaluation changed after taking the modules? I: What is your main objective when you evaluate students in the classroom?

I: Compared to the first round of observations before attending the online course, how did your ideas about how to include models in your classes change?

I: If you were able to do it, how did you assess students' understanding of basic ideas in science by using models?

I: Did you have the opportunity to evaluate the inquiry skills of the students in the classroom by having them think with models?

I: How did you evaluate the student models at the beginning, during and at the end of the unit? I: How did you translate the objectives of the curriculum into specific tasks to assess student models?

I: How did you use the evaluation to evaluate the coherence of the models built by the student?

I: How did you use evaluation to involve students in each phase of model generation, evaluation, and modification? Could you give me an example for each phase?

I: What criteria did you use to evaluate the students' models?

I: What assessment instruments did you develop to assess student models in the classroom and what was your purpose?

I: How did you communicate the formative feedback to the student after the generation, evaluation and modification of models?

I: What was your goal when you provided formative feedback on student models?

I: After a summative evaluation in which you involved the students to generate models, how did you provide feedback to the students?

I: How did you use the assessment results within the unit to adjust your pedagogy when engaging students to think with models?

I: Were you able to encourage your students to judge the quality of their role models or the claims of their classmates?

I: How did you use the assessment to help a student when they had trouble understanding a model? I: Could you describe how you gave the students opportunities to express their understanding of the model under study?

I: This is the last section of the interview, how did you organize the content in your classes within a unit after selecting the model you wanted to teach?

I: Did you incorporate the pre-existing ideas of the students when they thought with models?

I: How did you assess the students' progress in their understanding of a model?

I: Finally, how did you use assessment to help students enrich and refine their inquiry skills during model generation, assessment, and modification?

Appendix F: Factor Loadings for the Data Without Outliers

Item Factor Loadings for each Factor Solution for the QALMBT-Generic for the Data Without Outliers

Data without outliers			
Item	Factor 1	Factor 2	Factor 3
A12	**0.34**	0.26	0.08
A24	**0.65**	0.15	-0.05
A25	**0.56**	0.08	-0.02
A27	**0.51**	0.13	0.16
A28	**0.41**	0.02	0.04
A29	**0.55**	0.14	0.12
A30	**0.80**	0.02	-0.11
A31	**0.66**	0.11	0.02
A32	**0.51**	0.08	0.05
A33	**0.80**	-0.10	0.02
A34	**0.73**	-0.15	0.13
A1	0.05	0.28	**0.34**
A5	0.18	0.16	**0.32**
A6	-0.01	0.23	**0.46**
A11	0.08	0.24	**0.32**
A20	0.07	-0.07	**0.65**
A4	-0.14	**0.59**	0.19
A13	0.14	**0.58**	-0.05
A14	0.22	**0.35**	0.10
A17	0.19	**0.62**	0.01

Item Factor Loadings for each Factor Solution for the QALMBT-Modeling for the Data Without Outliers

Data without outliers			
Item	Factor 1	Factor 2	Factor 3
B10	**0.35**	0.25	0.26
B12	**0.40**	0.05	0.25
B15	**0.43**	0.19	0.23
B18	**0.41**	0.28	0.27
B19	**0.34**	-0.05	0.29
B20	**0.40**	0.19	0.11
B21	**0.70**	-0.04	0.14
B22	**0.51**	-0.09	0.19
B24	**0.73**	0.02	0.04
B25	**0.70**	-0.07	0.11
B26	**0.71**	0.01	0.15
B27	**0.62**	0.28	-0.10
B28	**0.74**	-0.14	-0.03
B29	**0.70**	0.20	-0.09
B30	**0.73**	0.09	-0.06
B31	**0.72**	0.20	-0.11
B32	**0.68**	-0.04	-0.01
B33	**0.80**	0.01	-0.02
B34	**0.86**	-0.16	0.03
B35	**0.52**	0.10	0.11
B1	0.10	**0.31**	0.07
B4	-0.02	**0.73**	0.10
B13	0.27	**0.46**	0.12
B17	0.24	**0.48**	0.20
B5	-0.02	0.04	**0.78**
B7	0.26	0.18	**0.38**
B8	0.06	0.04	**0.76**

Item Factor Loadings for each Factor Solution for the QALMBT-Epistemic for the Data Without Outliers

	Data without outliers		
Item	Factor 1	Factor 2	Factor 3
C2	**0.36**	0.22	-0.11
C7	**0.40**	-0.01	0.23
C8	**0.66**	0.03	-0.04
C12	**0.57**	-0.05	-0.03
C15	**0.43**	0.19	0.17
C16	**0.61**	0.00	0.18
C18	**0.59**	0.15	-0.09
C1	-0.07	**0.43**	0.01
C3	-0.04	**0.58**	0.16
C4	0.15	**0.55**	-0.02
C5	0.12	**0.50**	-0.08
C9	0.24	**0.37**	0.12
C6	-0.01	-0.03	**0.46**
C14	0.28	-0.06	**0.36**
C17	0.13	0.26	**0.35**
C19	-0.08	0.25	**0.59**
C20	0.26	-0.08	**0.53**

Note: Factor loadings in bold type were considered for the conceptual interpretation.

Appendix G: Full Version of the Rubric of Levels of Teacher Proficiency in Assessment Literacy in MBT (R-LPAL)

Disciplinary Knowledge	Novice	Advanced beginner	Competent	Advanced assessor
Knowledge of MBT as an instructional approach	Follows a lectured-based approach in which models are used to complement a definition and/or define components/features of a system (e.g., students observe a representation of an animal cell and the teacher defines the structures).	Uses models in the class mostly to present concepts, ideas, theories (e.g., the teacher presents and explain a curricular model, for example, a heart and lung diagram). It is different to novice in a way that the teacher explains a mechanism instead of merely emphasizing components/definitions.	Engages students in activities that involve the generation of a model; however, the generated model is barely used in the classroom.	Engages students in the generation of a model which is used as a research tool to generate information and understand a mechanism or phenomenon.
Explanatory and predictive power	The model is used only to present concepts or ideas that students study or memorize.	The activity or the model suggested by the teacher attempt to explain or represent a phenomenon.	The model is generated by students and it is used by them (students) to explain a phenomenon but not to make a prediction. If the role of predictions is mentioned but not observed in their pedagogy, teachers are labeled as competent.	Both explanatory and predictive power are tested by students after constructing or using a model.

Assessment purpose	Novice	Advanced beginner	Competent	Advanced assessor
Format of assessment	Implements assessment mainly in the same format (e.g., tests or resolution of exercises) in order to assess only factual or knowledge of a model.	Implements assessment in different formats but the focus is on assessing factual knowledge of a model rather than promoting modeling practices.	Implements assessment in different formats to assess students' reasoning by using models but the focus is still on teaching the curricular model (e.g., factual knowledge, mainly content).	Implements assessment in different formats to assess students' reasoning by using curricular models in order to promote the development of modeling practices.
Reasoning with a model	Asks students questions that do not challenge them to make judgements or explanations in science based on reasoning.	Ask students questions that are generic and does not challenge students to use their models to make judgements or explanations (e.g., what).	Asks students to make judgements by using a model that is facilitated by the teacher but the activities do not engage students in using their own models (e.g., how) (e.g., teacher asks students to conduct another investigation but models are not tested) .	Asks students to make judgements in science based on reasoning with their own models (e.g., why explanation; the model is tested in a new investigation).

Assess internal consistency of a model	Uses assessment to evaluate student understanding of a curricular model that has been taught in class (e.g., after a lecture)	Uses assessment to evaluate student understanding of a model constructed by a student which is aligned to the curricular model but the assessment does not involve the evaluation and modification of the initial model	Uses assessment to assess the consistency or coherence of a model generated and evaluated by a student but it does not involve the modification of the model	Uses assessment to assess the internal consistency or coherence of various models constructed by a student and engage students in a cycle of generation, evaluation and modification of the model

Scoring technique	Novice	Advanced beginner	Competent	Advanced assessor
Explanation of criteria	Offers an overall explanation of the purpose of the assessment without explaining each criterion that he/she will use to evaluate students' models even though s/he might do it from a generic approach of teaching.	Explains to students the criteria that he/she will use to evaluate students' models only for some of the tasks included in the summative assessment.	Explains to students the criteria that he/she will use to evaluate students' models for some of the tasks included in the summative and formative assessment.	After explaining the criteria, engages students in using the criteria to self-evaluate their own models.
Scoring tools	Always uses or suggests, for example, in an interview, the same scoring tool to evaluate the models generated by students (e.g., the items in a rubric do not change and are generic for each task; only assesses students with paper-and-pencil exams).	Uses more than one scoring tool (e.g., rubrics, checklists, standards) within a unit to evaluate a model generated by students; however, the scoring tool is generic and focuses on measuring students' understanding of core ideas rather than measuring modeling practices.	Uses different scoring tools (e.g., rubrics, checklists, standards) to evaluate the models generated by students but the scoring tools do not measure how students' reason with a model.	Uses different scoring tools (e.g., rubrics, checklists, standards) to evaluate the models generated by students and the development of modeling practices.

Communicate feedback	Novice	Advanced beginner	Competent	Advanced assessor
Formative feedback	Provides general feedback but it is not specific in terms of how to help students to evaluate their generated models.	Gives formative feedback to students about their generated models only for some of the tasks that involve modeling.	Gives formative feedback to students about their generated models for the majority of the tasks that involve modeling but the feedback does not always support for refinement of students' initial models.	Gives formative feedback to students about their generated models for each of the tasks that involve modeling which support students' achievement of the curricular model.

Results of assessment	Communicates the results of the assessment for the whole assessment (e.g., the grades in an exam) without breaking down each item or question.	Communicates the results of the assessment in order to inform the correct answers but he/she does not focus on clarifying students' alternative ideas about the model.	Communicates the results of the assessment (e.g., summative) only for some of the activities that involve thinking with models in order to help each student to achieve a better understanding of the expected model.	Communicates the results of the assessment (e.g., summative) for each task in order to help each student to achieve a better understanding of the expected model.
Consensus model/explanation	Asks students to repeat the main ideas taught in class by the teacher (e.g., about a curricular models) in order to identify their understanding.	Asks students to discuss about their explanations, for example, in small groups, but he/he does not challenge students to create a consensus explanation.	Promotes the generation of a consensus explanation only in a small group of students but the explanations/model/claims are not discussed later in the class.	Promotes the generation of a consensus explanation that helps students have a similar understanding of the phenomenon under study. The whole class participates in the generation of the explanation.

Interpretation of Assessment	Novice	Advanced beginner	Competent	Advanced assessor
Judge students' understanding	Uses assessment to measure students' understanding of the curricular model but he/she does not interpret the assessment to clarify students' ideas or adjust his/her instruction.	Uses assessment to measure students' understanding of the curricular model and inconsistently includes students' ideas from the interpretation of assessment.	Uses assessment to judge students' understanding about the phenomenon to be modeled and clarifies some ideas about the model.	Uses assessment to judge students' understanding about the phenomenon to be modeled and adjusts his/her instruction based on the results obtained from the assessment.
Refinement of a model	Uses assessment results with the academic purpose, for example, to assign grades to students but he/she does not use the information to help students refine their ideas or models.	Uses assessment results to measure students' current understanding about a model but does not consistently use the information to help students to refine their ideas or models.	Uses assessment results to identify the most common ideas about the curricular model that students have reshaped in order to evidence whether students are reaching the expected curricular model.	Uses assessment results to compare how students' ideas about the model studied have been reshaped, for example, within a unit before reaching the expected curricular model.
Missing elements in a model and adjustment of the instruction	Uses assessment to identify students' understanding about a model but s/he does not explore the missing elements that students have not understood or	Uses assessment to locate evidence about the missing elements that students have not understood regarding the model but he/she does not develop new tasks to help students	Uses assessment to locate evidence about the missing elements that students have not understood regarding the model under study and indicates the elements that students did not include.	Uses assessment to locate evidence about the missing elements that students have not understood regarding the model under study and adjust his/her pedagogy to coach his/her students during the

	incorporated in their models.	revise and modify their models.		revision and modification of a model.

Engage students in assessment	Novice	Advanced beginner	Competent	Advanced assessor
Construction of Assessment criteria	Encourages students to generate and share an explanation to their classmates (e.g., how to solve a problem, define a concept) in order to help them assess their own understanding; however, students are not challenged to develop or use assessment criteria in order to comment on their classmates' ideas/models.	Facilitates the assessment criteria that students should use to assess their peers but the results of the assessment are not shared among students (e.g., a student only assign a score but does not discuss or provide feedback) .	Challenge students to develop or use assessment criteria to evaluate the models constructed by their classmates and asks students but there is not a reflection that allows the modification or refinement of the original model.	Challenges students develop or use assessment criteria to evaluate the models constructed by their classmates in order to encourage others to reflect about epistemic criteria for good models (e.g., regarding the nature and purpose of a model, scope of a model, limitations, etc).
Evaluate new information	Shows a model and explains the variables included, for example in a diagram. S/he provides information (e.g., a table with boiling points) that is used to understand a model. The teacher uses this information to evaluate the utility of a model instead of challenging students to do it.	Asks students to generate a model and then asks them to review information without challenging students to evaluate their models.	Asks students to analyze and evaluate new information by using their models, but students are not challenged to modify their models. Rather, they only assess the utility of their models to explain a phenomenon.	Asks students to analyze new information to promote the evaluation and modification of models that help them collect evidence to show the utility and explanatory and predictive power of their generated models.

Knowledge of assessment of ethics	Novice	Advanced beginner	Competent	Advanced assessor

Feedback for each student	Only reinforces students correct answers and does not use students' wrong or incomplete answers to identify if students are having difficulties about a model or their generated models.	Offers explicit guidelines to the class in order to improve students' understanding of the curricular model but he/she does not provide individual feedback.	Provides feedback only for some students in order to clarify their understanding of the curricular model and he/she does not inform or summarize the main points to the rest of the class.	Provides feedback for each student in order to ensure that each of them achieves the same understanding of the curricular model. For example. s/he summarizes the main points to ensure that each student achieve a similar understanding of the curricular model.
Use of the results of the assessment	Uses assessment only to grade students performance instead of using it as a tool to enrich students' understanding about a model and give them the opportunity to modify their models.	Communicates and uses the results of the assessment in order to inform the correct answers or teach the expected curricular model.	Communicates and uses the results of the assessment in order to inform the correct answers and also provides some guidelines to help students revise their models and understanding.	Communicates and uses the results of the assessment (e.g., summative) to help each student to achieve a better understanding of the expected model and design specific activities to help student reshape their ideas before moving to the next core ideas that requires a good understanding of the core idea recently assessed.

Scaffolding and Learning progression	Novice	Advanced beginner	Competent	Advanced assessor
Scaffolding activities	Incorporates activities that always measure the same skills (e.g., rote learning).	Incorporates activities with a similar level of complexity but in some occasions he/she challenges students to reach higher skills (e.g., students elaborate claims, use evidence, use a model).	Incorporates activities that vary in complexity which assess students' understanding of the model; however, the focus is not always on enriching or modifying the model.	Incorporates scaffolding activities or tasks which progress in complexity in order to assess students' understanding of the model and encourage them to evaluate and modify their models.
Adaptation of the curricular model	Is not aware of the difficulties that students have to understand the system under study or the curricular model and he/she does not adapt his/her pedagogy to reduce the	Adjusts the complexity of the curricular model to facilitate student' understanding of the system under study; however, the model lacks of some important components related to modeling practices	Adjusts the complexity of the curricular model to facilitate student' understanding of the system under study but he/she does not lead students in conversations to enrich and refine their ideas about the model.	Adjusts the complexity of the curricular model to facilitate student' understanding of the system under study and leads students in conversations to enrich and refine their ideas about the model that should study according to the provincial science curriculum (e.g., attributes/ generalizability and limitations of a model;

	complexity of the curricular model.	(e.g., generation phase) .		suggests new variables or factors that students could incorporate in their models; anticipates elements of a model that may be challenging for students) .

Appendix H: Data from the Validation of the QALMBT (English Version)

Ratings on the QALMBT-Generic

QALMBT-Generic										
	ENGLISH							SPANISH		
Item	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	% agreement		Expert 1	Expert 2	% agreement
1	X	X	X	X	X	1		X		0.5
2	X		X	X	X	0.8		X	X	1
3				X	X	0.4		X	X	1
4			X	X	X	0.6		X		0.5
5	X		X	X	X	0.8		X	X	1
6	X		X	X	X	0.8		X	X	1
7	X	X	X	X	X	1		X	X	1
8	X	X	X	X	X	1		X	X	1
9	X	X	X	X	X	1		X	X	1
10	X		X	X	X	0.8		X	X	1
11	X	X	X		X	0.8		X		0.5
12	X	X	X	X	X	1		X	X	1
13	X		X	X	X	0.8		X		0.5
14	X	X	X	X	X	1		X	X	1
15	X		X		X	0.6		X	X	1
16	X	X	X		X	0.8		X	X	1
17	X	X	X	X	X	1		X	X	1
18	X	X	X	X	X	1		X	X	1
19	X		X		X	0.6		X	X	1
20	X		X	X	X	0.8		X	X	1
21	X		X	X	X	0.8		X	X	1
22	X	X	X	X	X	1		X	X	1
23	X	X	X	X	X	1		X	X	1
24		X	X	X	X	1		X	X	1
25	X	X	X	X	X	1		X	X	1
26	X	X	X	X	X	1		X	X	1
27	X	X		X	X	0.8			X	0.5
28	X	X	X	X	X	1		X	X	1
29	X	X	X	X	X	1		X	X	1
30	X	X	X	X	X	1		X	X	1
31		X	X	X	X	0.8		X	X	1
32	X	X	X	X	X	1		X	X	1
33	X	X	X	X	X	1		X	X	1
34	X	X	X	X	X	1		X	X	1
35	X	X	X	X	X	1		X	X	1
36			X		X	0.4				
37			X	X	X	0.6				

Note: Items rated as 1 (essential) are indicated with an X. Items 36 ("I use assessments to evaluate students' creativity in the science classroom") and 37 ("I provide feedback for each student to ensure that each of them achieves the same understanding of the core ideas") were eliminated from the final version of the questionnaire.

Validation QALMBT-Modeling (English version)

Ratings on the QALMBT-Modeling

QALMBT-Modeling										
ENGLISH								SPANISH		
Item	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	% agreement		Expert 1	Expert 2	% agreement
1	X	X	X	X	X	1		X		0.5
2	X	X	X	X	X	1		X	X	1
3			X	X	X	0.6		X	X	1
4	X		X	X	X	0.8		X		0.5
5	X		X	X	X	0.8		X	X	1
6	X	X	X	X	X	1		X	X	1
7	X	X	X	X	X	1		X		0.5
8	X	X	X	X	X	1		X	X	1
9	X	X	X	X	X	1		X	X	1
10	X	X	X		X	0.8		X		0.5
11	X	X	X		X	0.8		X	X	1
12	X	X	X	X	X	1		X	X	1
13	X		X	X	X	0.8		X		0.5
14	X	X	X	X	X	1		X	X	1
15	X	X	X	X	X	1		X		0.5
16	X	X	X	X	X	1		X	X	1
17	X	X	X	X	X	1		X	X	1
18	X	X	X	X	X	1		X	X	1
19	X		X		X	0.6		X	X	1
20	X		X	X	X	0.8		X	X	1
21	X	X	X	X	X	1		X		0.5
22	X	X	X	X	X	1		X		0.5
23	X	X	X	X	X	1		X	X	1
24		X	X	X	X	0.8		X		0.5
25	X	X	X	X	X	1		X	X	1
26	X	X	X	X	X	1		X		0.5
27	X	X	X	X	X	1		X	X	1
28	X	X	X	X	X	1		X		0.5
29	X	X	X	X	X	1		X	X	1
30	X	X	X	X	X	1		X	X	1
31		X	X	X	X	0.8		X	X	1
32	X	X	X	X		0.8		X	X	1
33	X	X	X	X	X	1		X		0.5
34	X	X	X	X	X	1		X	X	1
35	X	X	X	X	X	1		X	X	1
36			X		X	0.4				
37			X	X	X	0.6				

Note: Items rated as 1 (essential) are indicated with an X. Items 36 and 37 were eliminated.

Validation of the QALMBT-Epistemic (English version)

Ratings on the QALMBT-Epistemic

QALMBT-Epistemic										
ENGLISH								SPANISH		
Item	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	% agreement		Expert 1	Expert 2	% agreement
1	X	X	X	X	X	1		X	X	1
2	X	X	X	X	X	1		X	X	1
3	X	X	X	X	X	1		X	X	1
4	X	X	X	X	X	1		X	X	1
5	X	X	X		X	0.8		X	X	1
6	X	X	X	X	X	1		X	X	1
7	X	X	X	X	X	1		X	X	1
8	X	X	X	X	X	1		X	X	1
9	X	X	X	X	X	1		X	X	1
10	X	X	X		X	0.8		X		0.5
11	X	X	X	X	X	1		X	X	1
12	X	X	X	X	X	1		X	X	1
13	X	X	X	X	X	1		X	X	1
14	X	X	X	X	X	1		X	X	1
15	X	X	X	X	X	1		X	X	1
16	X		X	X	X	0.8		X	X	1
17	X	X	X	X	X	1		X	X	1
18	X	X	X	X	X	1		X	X	1
19	X	X	X	X	X	1		X	X	1
20	X	X	X	X	X	1		X	X	1