INVESTIGATING SUBSTANTIVE VALIDITY EVIDENCE USING ACTION-PROJECT METHOD FOR THE GOAL ATTAINMENT SCALING MEASURE

by

Sneha Shankar

BSc., University of Waterloo, 2003
MSc., University of Waterloo, 2006
MOT, University of British Columbia, 2008

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Measurement, Evaluation and Research Methodology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

September 2019

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

Investigating Substantive Validity Evidence Using Action-Project Method for the Goal Attainment Scaling Measure

submitted by Sneha Shankar in partial fulfillment of the requirements for

the degree of Doctor of Philosophy

in Measurement, Evaluation and Research Methodology

Examining Committee:

Sheila K. Marshall, Social Work
Supervisor

Bruno D. Zumbo, Measurement, Evaluation and Research Methodology
Supervisory Committee Member

Richard A. Young, Counselling Psychology
Supervisory Committee Member

Sterett H. Mercer, Special Education
University Examiner

Chris Lovato, School of Population and Public Health
University Examiner

Abstract

One aspect of validation is response processes, which examine individuals' interactions with a measure. Current research on response processes tends to focus on tool interpretations as a function of one user. It is unclear how validation processes consider another user who may also be involved with use of a tool. My dissertation applies the Action-Project Method (APM) to investigate the way two people engage with the Goal Attainment Scaling (GAS) measure. Using APM, this dissertation captures an array of response process information and delves into the goal construct that is evaluated by the GAS.

Chapter One introduces the main components of this dissertation: validity, APM, and GAS. I explain why APM is suitable to study response processes, and how it will provide access to the joint processes of individuals using the GAS.

Chapter Two starts by exploring how validity evidence has been gathered for the GAS. I show the ways researchers have investigated validity for a measure used between two people and highlight gaps in evidence. I discover that a clear goal construct has not been identified as most validity evidence reports on relations of the GAS to other variables. This review demonstrates that validity evidence does not consider the influence of theory or response processes, which limits the inferences that can be made about the GAS.

Chapter Three aims to bridge the identified gap in validity evidence using an innovative method to explore response processes. APM is used to examine how and why a sample of therapists engage with the GAS. The complexity of interacting with the GAS is unveiled, as well as the assumption that the construct underlying the GAS is dually goal intentions and goal attainment.

The final Chapter describes how the concepts and methods come together, novel contributions, limitations, and implications for researchers. Overall, this dissertation advances validity research through its application of a novel method to investigate response processes. This research also expands conceptions of validation to include investigation of joint processes and demonstrates how response process data can go beyond cognitive processes to include actions, emotions and motivation.

Lay Summary

My dissertation contains two related studies to investigate how a measure is used between two individuals, and applies an innovative method to study this interaction. The first study reviews the literature to understand how a measure, the goal attainment scaling, has been evaluated in terms of the goal concept it claims to assess, and how the interaction between two individuals has been investigated. The second study uses action-project method to investigate joint engagement with goal attainment scaling. The results from the first study reveal gaps in the way evidence for validation is collected. The findings from the second study demonstrates that action-project method is an exceptional method to observe how people interact with measures and is able to extract information that goes beyond traditional investigations of cognitive processes to include actions, emotions and motivation. This dissertation contributes a novel method to help steer the future of validity research.

Preface

Chapter Two. A version of chapter two of this dissertation has been published in the Journal of Psychoeducational Assessment, with the title, '*A systematic review of validation practices for the Goal Attainment Scaling measure*.' My supervisor, Dr. Sheila K. Marshall and co-author/committee member Dr. Bruno D. Zumbo, both at the University of British Columbia helped me develop the research question and also helped me with revision requests by the editor and reviewers. My supervisor, Dr. Sheila K. Marshall, was the second reviewer in the selection of articles for inclusion and in this role she provided a secondary perspective to verify my coding sheet that was used to screen and select final articles. I wrote the entire manuscript, determined the search criteria, coding for analyses, compiled all results and completed revisions before publication. The following figures and tables are taken from this manuscript: Tables 2.1, 2.2, 2.3 and Figure 2.1

Chapter Three. Approval was obtained by the behavioural research ethics board at UBC (Ethics approval: H18-01915) for the protocol used in the study described in Chapter 3 of the dissertation. I conducted all interviews in this study.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

APM - Action-Project Method

CAT- Contextual Action Theory

GAS - Goal Attainment Scaling

Acknowledgements

I have thought a fair bit about what to include in this section. It is after-all forever imprinted and I hope will resound with me in the years to come. As there are too many things to say to capture the entirety of my PhD experience, I highlight some memorable people and moments—in both fun and emotive expressions. It is a simple expression of myself.

I start by thanking my dear friend and sister, Angela Parker-Jervis whom I have had the good fortune of knowing for many years. I am so grateful to have cultivated an additional family on the West Coast and feel so lucky to have you, John, Alessia and Aiden. You have witnessed all of this journey and its preamble—you even saw the questions that led here. I thank you for being my ally throughout this journey, listening to many PhD musings and sharing with me in moments of joy, sadness and of course laughter. As true friendships go, sharing a coveted treat with me during my most downtrodden moment was a turning point. Even when I initially declined your offer, you mindfully put it aside knowing such a refusal would be foolish. This moment inspired me to write an ode that playfully captures the absurdity of my PhD journey. This poem is an homage to the spectrum of emotions I felt during this degree, and I hope it brings a smile to those who read it. It is a metaphorical poem about an Italian pastry:

Zeppole, it was a moment in time
An indulgence as you melted so slow
Pastry and tasty sugar together so sublime.
I wish I knew what I now know
You were light, you were sweet, you created a life shift.
Honey emerged, an unfamiliar treat.
Fluffy dough fried to perfection, and no excess grease
I didn't expect this level of uplift
Not burnt in oil, sugary residue sweet.
A dessert that was simply, a masterpiece.

Zeppole, an incredible delight
I will not soon forget
how you helped me write
To focus, reset, and stop the fret
Of all the sweets I adore
I never expected influence from you
A bread like relief
Not something to ignore
I thank you, merci beaucoup
You truly helped my belief.

To all the others that have been a part of this dissertation, I thank you for your positivity, innumerable embraces and continued encouragement. I send a very special thanks to the following friends who played an important role in this achievement. Kerstin Crosthwaite-Daudrich, your funny and uplifting messages were always appreciated. You always remind me to ask "...but how can we do better?" To Tobin Copley, who has been a pillar of support in this academic pursuit. Thank you for your generous dinners and sound advice whenever needed.

Melanie Power, hearing about evidence from your law perspective helped me pay attention to words, both written and spoken. To Ilana Waniuk, it is an incredible honour to grow up together. I always relish the creative energy you bring, and your encouragement to use mine. Kelly Skinner and Claire Davies, it has been lovely to reconnect and collaborate. Thank you for all your advice, attentive ears and incredible support. I am lucky to have researchers with strong principles to both influence and inspire me. To my family – the knowledge and strength that I needed was passed down by you. I thank you. There are many others that have been there for important moments, such as: Ashley Pullman, Regina Casey, Ann Webborn and Shelagh Smith.

Importantly, I am indebted to all the individuals who helped me with my study as none of this would be possible without you – thank you! Dan Ji provided much help with testing and analysis. Thank you for sharing your insights and being an excellent officemate!

Ultimately, this PhD is a reflection of my work environment and all the individuals who believed in me and stood by me. This includes MERM colleagues who I have been so fortunate to know. I thank you for our many authentic conversations and inspiring me with your fortitude and tenacity to move forward. I send a special thanks to the following: Oscar Astivia, I thank you for sharing with me in critical thinking, debating concepts in measurement and most of all engaging in principled measurement; Keren Roded, our lunches, genuine heartfelt conversations and runs were always so nourishing. I thank you for your kindness, encouragement, and always reminding me what it takes to be a strong woman. Michelle Chen, I always love your wise, thoughtful perspectives and Karen Yan, you were pivotal in helping me navigate the system.

To my supervisors, Sheila, Bruno and Richard, I thank you for all of your teachings—helping to mold my learning, cultivate my thinking and mentor me. Most of all, I thank you for giving me the opportunity to flourish. Richard, I am so fortunate that our paths have crossed. I enjoyed all the conversations we have had about methods and philosophy, and how my perspectives of action have expanded. Bruno, your enthusiasm for research and teaching is so contagious. Great teachers make difficult concepts accessible to a wide audience, and not only did you help me access them, you also guided me to excel. Sharing spaces in several of your classes has been an immense joy and honour, and an absolute highlight of my degree. Sheila, this PhD has taught me many things, but most of all it has taught me about shared humanity in a very privileged setting and how simple gestures can mean so much. My most memorable encounter at UBC has been working with you. You taught me more about the breadth of measurement then I could have imagined. In particular, you contributed to a number of my learnings, such as teaching me how to work with applied data and seeing the imperfections in data but remaining both accountable and principled. All the while, you showed me how integrity, respect and kindness go so far. You always believed in me and for that my deepest and sincerest gratitude extends to you, as I would not be here without you. Thank you for showing me that I do matter.

~

A note to my future self: I hope to look back at this document and be reminded of resilience, goodness and the moral obligations that come with knowledge. This PhD experience has given me the opportunity and privilege to think and allow my thoughts to develop and grow. It is this wisdom of thinking that makes me so very thankful. I am grateful I have lived this 'what if'!

Dedication

For those who are exceptions to the rule, the ones who do not fit into the odds and the outliers.
And those who stand with them.

Chapter 1. A brief background

1.1 Illustrating the problem

Compare these two scenarios:

Scenario #1: You are a teacher preparing to use a measurement tool[1]*with students in your class. You choose to use this tool to help with your evaluations of each student's progress later in the school year. Your decision to use this tool is based on the understanding that there is good evidence to support its use (e.g. validity and reliability evidence).*

Scenario #2: You are a teacher preparing to use a different measurement tool. In this instance, you are sitting face-to-face with a student in your class preparing to use this tool. You complete the tool with each student. Similar to the previous situation, you would like to use this tool to help with your evaluations of the student's progress later in the school year, and again see that psychometric evidence has been investigated. You have been trained to use this tool and have the expertise to administer the tool. During each student interview, you try and use the tool in the same way, but you recognize that students are not just responding to the tool, they are also responding to you. As you are both working towards completing the tool, students are making interpretations about the tool, as well as your administration of the tool. All this makes you realize that there are essentially two people involved with this tool at the same time - yourself and the student. You wonder how this dyadic engagement between yourself and the student has been considered during psychometric testing of the tool.

[1] In this dissertation the term "tool" will be used interchangeably with the terms "test" or "scale" or "measure" to indicate a measurement process that is used to identify differences between individuals and make inferences about a person's behaviour.

The Problem: Current validity information for tools tend to focus on tool interpretations as a function of one user (i.e. Scenario 1). When two people are involved with use of a tool, as in Scenario 2, it is unclear how validation processes consider another user who may be jointly using the tool. In particular, validity evidence does not appear to account for a teacher-student dyad (for instance) when validity information is obtained for relevant measurement tools. It seems that the teacher is presumed not to be influential in the operation of the tool or subsequent responses. Thus, two questions that are answered in this dissertation include: (1) How are tools evaluated when used in a joint manner? (2) How do psychometricians evaluate substantive validity evidence, such as response processes in these situations?

My PhD dissertation tackles this problem and evaluates validity evidence based on response processes by using a method called Action-Project Method, with a measurement tool called Goal Attainment Scaling. This introductory chapter provides an outline to the components of this dissertation with the remainder of the dissertation attempting to address this psychometric problem. Starting with an overview of the measurement concepts of validity and validation, this chapter then presents the Action-Project Method as a viable method to investigate response processes. Finally, Goal Attainment Scaling, which acts as a model in this dissertation for any tool whereby two users are engaged with a measure, is described.

1.2 The meaning of validity and validation

Validity is a cornerstone of measurement. According to the *Standards for Educational and Psychological Testing*[2], validity is "the most fundamental consideration in developing tests and evaluating tests" (p. 11) and refers to the degree to which evidence supports the intended interpretations for a proposed test use (American Educational Research Association [AERA],

[2] Henceforth termed the *Standards*

American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). Validity is about the "inferences, interpretations, actions, or decisions" that are based on test scores (Hubley & Zumbo, 2011, p. 220) and provides meaning to a test through the interpretations that are made from test scores (Messick, 1975).

Gathering information about a test's meaning entails determining the extent (or degree) to which various strands of evidence support a particular inference, and also establishing alternative inferences that are less well supported (AERA et al., 2014; Messick, 1989a, 1989b, 1995). Establishing the evidential basis to infer a measure's meaning centers on the concept of construct validity (Messick, 1975). The term "construct" refers to an unobservable phenomenon or behaviour, which is assumed to capture the entity one aims to measure; and in the unified (or construct) view of validity, the construct is the overarching component (Downing, 2003; Messick, 1989b). In this view of validity, various sources of validity evidence are integrated. These sources of evidence include: test content (relevance and representativeness of test items), response processes (interaction between test user and test items), relations with other variables (e.g. convergent and discriminant validity), internal structure (dimensionality), and consequences of testing (intended, unintended uses and misuse) (AERA et al., 2014). These five sources help establish a nomological network or a "chain of inference" (Cronbach & Meehl, 1955, p. 291) that links validity evidence to the test construct, which come together around the construct as, "the whole of validity" (Loevinger, 1957, p.636). According to the *Standards*, a sound validity argument rests on a summary judgement that integrates evidence to support the interpretations of test scores for a specific purpose (AERA et al., 2014). Since interpretations include meaning or explanation, interpreting a score is analogous to explaining the meaning of the test score (Kane, 1990).

The explanations that are incorporated in our interpretations may be based on laws, theories, or assumptions about the relationship between test scores and the construct (Kane, 1990). In order for construct validation to be approximated, an explanation or theory is needed to encompass all the evidence (Loevinger, 1957). Indeed, strong forms of construct validity evidence includes explanatory interpretations that are theory-based, whereas weak forms are characterized by correlations of the test with other variables (Cronbach & Meehl, 1955; Kane, 2001; Zumbo, 2009). Theory has explanatory power for the variations that are observed among test scores, and explanations serve as a regulative ideal to guide inferences (Zumbo, 2007a, 2009). Thus, questions about the manner in which evidence is interpreted and justified, as in the example presented at the outset of this dissertation, are questions about test validity that require explanations. These explanations help to understand why things are as we find them (Zumbo, 2009) and what evidence enables a test to be used in the manner that is suggested (Messick, 1989b).

To determine explanations, this dissertation employs the unified view of validity as originally described by Cronbach and Meehl in 1955 and currently endorsed by the *Standards* (AERA et al., 2014). The unified view will be used to guide an examination of test validity for measures that can be used jointly, whereby joint refers to processes that occur both at the individual and group level (Valach & Young, 2002). As validity evidence provides the justifications or explanations for variations in test scores, the process of validation must serve this definition (Zumbo, 2009). Thus, in addition to the unified perspective of validity, this dissertation uses Zumbo's (2009, 2017a) explanation-focused approach to validation to explore engagement with a test that involves a dyad. An explanatory approach takes an ecological view of item responding to motivate a focus on the contextual factors or the environment surrounding

a test, such as the social context (Zumbo, 2009, 2017a). The basic idea underlying the explanatory approach involves understanding the reasons, "why an individual responded in a particular way to an item or scored a certain value on a scale" (Zumbo, 2009, p. 70-71), and this logic can be used to help bridge the inferential gap between constructs and test scores for measures that are used jointly.

1.3 The need for substantive evidence

Understanding the process by which individuals arrive at their answers on a measure means investigating the substantive component of validity (Borsboom, Mellenbergh, & van Heerden, 2004; Loevinger, 1957; Messick, 1975; Zumbo & Hubley, 2017). Loevinger (1957) first introduced the substantive component of validity, which is defined as the degree to which the content of the test is accounted for in terms of the construct that is being measured and the context of measurement (with "context" including both theory and test behaviour). In the unified view of validity, test responses or test behaviours are espoused within validity evidence pertaining to response processes[3]. As such, substantive validity evidence has some overlap with responses processes, as well as similarities with validity evidence pertaining to test content, which are two sources that are described in the *Standards* (AERA et al., 2014). Validity evidence pertaining to test content analyzes the relationship between the content of the test and the construct measured, and response processes examines how individuals interact with a test and test items (AERA et al., 2014). Together, these sources evaluate the *representativeness* of a measure, or more specifically, how representative a measure is towards the construct it aims to measure and also the degree to which processes that are involved represent the construct that is being measured (Loevinger, 1957; Messick, 1995). Validity evidence pertaining to test content

[3] In this dissertation, the terms 'test behaviour' and 'response processes' both pertain to how individuals interact with a measure.

and response processes are complementary since they both address how adequately the content on the test represents the content domain and also how the content of a test may be interpreted differently across respondents (Padilla & Benítez, 2014). Since the goal of construct validation is determining the interpretability or meaningfulness of test scores, substantive evidence provides critical information to contribute towards this evaluation.

The substantive component of validity includes a specific examination of theory and response processes, which provide empirical evidence that a measure is consistent with its intended interpretation (Loevinger, 1957; Messick, 1980). Theoretical rationales support the interpretations and actions that are based on test scores (Messick, 1989b), and even if the interpretations are more practical then theoretic, theory has a role in the validity argument (Kane, 1990). Examining the theoretical rationales behind the construct provides a glance into assumptions about a measure and helps to define the boundaries of a construct (Kane, 2001). Likewise, response processes can provide data that inform how interactions during a test can influence test performance (Hubley, 2017). Considering the context of measurement, such as practitioner or educator knowledge and participant responses, enables a direct look into how a specific construct is being measured, as well as assumptions that may surround this construct.

Since substantive validity evidence provides information that relate to observed consistencies in test responses and the extent to which the content on the test represents the construct intended (Loevinger, 1957; Messick, 1980; Messick, 1995), the substantive component of validity can provide explanations related to dyadic engagement during use of a measure. By considering the substantive component of validity, this dissertation examines the context of measurement and the interpretability of a test. In particular, response processes are examined in a dyadic context to develop explanations for engagement with a joint measure.

1.4 Using APM to investigate substantive evidence

The context of measurement is an important consideration for how test responses are formed. In addition to theory, understanding response processes and how a construct is being measured means evaluating more than verbal or written responses and the content on the test (Loevinger, 1957). Yet, the most researched response process are cognitive processes, which are typically examined through verbal strategies such as think aloud protocols or cognitive interviewing (Launeanu & Hubley, 2017). Since affective and emotional processes can also influence test performance (Launeanu & Hubley, 2017; Leighton, 2015), response process methods need to go beyond cognitive processes.

In addition, methods that investigate response processes tend to focus on validity evidence for one respondent. Common methods such as think aloud protocols, cognitive interviewing, or eye tracking are methods for investigating response processes which assume the respondent is the only user of the test. Therefore, methods to investigate response processes are restricted in their abilities to capture joint response process. Joint processes include individual and group goals and actions (Valach & Young, 2002), which would occur in response to a test. As this dissertation investigates validity evidence when two individuals engage with a test, it is necessary to first conceptualize dyadic actions (e.g. teacher-student), and also consider how their joint actions construct meaning in response to a test. As such, the method to investigate validity evidence based on response processes needs to consider dyadic engagement. One approach that is suitable for dyadic or triadic contexts and offers a way to examine response processes for test engagement involving more than one person, is Action-Project Method (APM; Marshall, Zaidman-Zait, Domene, & Young, 2012; Young, Valach, & Domene, 2005).

APM is conceptually grounded in action theory, which understands action as inherently goal-directed and purposive although not necessarily rational (von Cranach, Kalbermatten, Indermuhle, & Gugler, 1982). APM draws on contextual action theory (CAT; Young, Valach, & Collin, 2002), which provides a framework to support relational processes and joint actions (Young et al., 2001). CAT considers action from three perspectives (Domene, Valach, & Young, 2015): (a) manifest behaviour – which is the observable behaviour necessary to carry out an action, (b) internal processes – the cognitive and emotional processes experienced during an activity and (c) social meaning – which are the explanations about one's actions that also consider the intentions and purpose of the action. As contextual action theory considers action as inseparable from and integrative of: cognitive processes, emotion, psychological processes, social meaning and one's intentional stance (Valach, Young, & Domene, 2015). An advantage of CAT is its applicability towards actions that are intentional and nonlinear, as well as its consideration of levels of analyses above the individual (joint). Thus, APM provides a way of obtaining explanations, in order to understand why individuals respond the way they do. As response processes have not been investigated in dyadic contexts, APM will be used as a method to investigate response processes for one respondent in this context.

1.5 The GAS as a tool to investigate substantive validity

To delve into the process of validation and understand how two or more participants jointly engage with a tool, the Goal Attainment Scaling (GAS; Kiresuk & Sherman, 1968) acts as a model for experimentation in this dissertation. The GAS has been in use for over 50 years and it is widely used internationally and across disciplines. This tool provides an exemplary model for tackling the psychometric problem in this dissertation, since the GAS involves more than one user and the content in the tool is formed by user(s) during engagement with the tool. The

content in the GAS are goals, which are jointly constructed and pertain to a specific context, such as a classroom or therapeutic setting. As the GAS is widely used, investigating substantive validity evidence for this measure will help to shed light on the problem outlined at the outset: how are tools evaluated for response process evidence when used in a joint manner?

1.5.1 About the GAS

Originally developed for use in mental health settings, the GAS is a tool in which users set, scale and score relevant goals for treatment or intervention outcomes. The GAS was originally used with a committee (or "goal selector"), who determined a set of realistic goals for a patient or client; this same committee then graded and scaled goals according to likely treatment outcomes (Kiresuk & Sherman, 1968, p. 445). In later descriptions of the GAS, Kiresuk et al., (1994) explain that goals can be set by a therapist or intake worker, client, or both the therapist and client. It is considered to be an individualized criterion measure that evaluates "intended change that will be pursued in treatment" (Kiresuk, Smith, & Cardillo, 1994, p. 2).

The GAS has been described as having "no fixed content" (Kiresuk et al., 1994, p. 167) and uses a combination of approaches, both qualitative and quantitative, to develop and then scale goals. The goals that form the content of the measure are concerned with behaviours, symptoms or characteristics that a treatment or intervention is attempting to change (Kiresuk et al., 1994). Unlike standard tools which use an already prepared self-report format, the content on the GAS is formed by the respondents and/or users during their use of the tool. Once goals are set, these goals are scaled by individuals who have knowledge of the treatment area in order to measure its effectiveness. It is assumed that users of the GAS will bring relevant, prior knowledge of the program or treatment to determine what goals are realistic for the client or student, and to grade and scale the goals appropriately. Prior knowledge of the context in which

the tool is used is not included in the measure in the form of items, but it is needed for proper use of the measure. As such, the tool can be used in educational and clinical settings but goals determined for the GAS will be different across contexts. Thus, the content of the GAS is determined, in part, by the users of the tool and how they apply prior knowledge (e.g. from education or experience) while using the tool. Ultimately, using the GAS means assessing a client's skill level in a particular problem area (for example), developing and scaling a goal to reflect a client's achievement that will be the result of the treatment or intervention, and identifying variations in goal attainment on a scale that indicates potential movement above or below treatment expectations. To outline how goals are set and scaled, Kiresuk et al. (1994) outline a step-by-step process as an example of their use in a mental health setting (see Appendix A).

The GAS is typically operationalized in an interview format to determine goals. Interviews involve interactions between users, such as a therapist and client or teacher and student. The users generate information about the client's specific problem areas and set goals. As goals are set, a rating scale is used to represent different levels of the goals. The rating scale is used by test evaluators, such as the clinician or teacher to judge expectations and performance. To determine the scale points and scale range, one user, typically a clinician or teacher, outlines varying levels of change for each goal(s). The suggested scale range has anchors at 1-point intervals, typically from -2 to +2. The determined goal is scaled at the expected level of outcome or level 0, with scale points above and below reflecting the change from treatment at follow-up. As indicators or items represent the construct (Pedhazur & Schmelkin, 1991), goals act as a representation of an indicator on the GAS, where the indicators collectively represent the goal construct. Furthermore, each goal level acts as a checklist of sorts, since reaching points above or

below expected outcomes indicate change in behaviour as a result of treatment (Kiresuk et al., 1994). All goals that are developed using the completion of the GAS are measured on the same scale and Appendix B is an example of the GAS format along with a sample goal (Kiresuk et al., 1994).

Although the developers of the GAS stipulate that only two scale points need to be specified, they also explain that a variety of scale points can be used. Once scaling is complete, the individuals administering the GAS determines a set of weights for the goals to reflect the value and successful appraisal of each goal, a composite score is determined and transformed into a standardized goal attainment T-score (Kiresuk et al., 1994). This process of scaling goals and calculating T-scores act to quantify one's goals and aims to evaluate the effectiveness of a treatment program or intervention and assess the "degree of change" (Kiresuk et al., 1994, p. 5).

1.5.2 Evaluating substantive validity evidence for the GAS

Evaluating validity evidence for the GAS means evaluating the overall construct validity of the measure. In the process of validation, the evidence obtained must relate to the construct that is being measured in the GAS, which in this case is the goal construct. Kiresuk et al. (1994) draw attention to the importance of construct validation, and they also articulate reasons various types of evaluation are limited by the GAS format. They argue the mean or standard T-score generated by the GAS score, provides information about validity and can be directly interpreted to evaluate change given the following assumptions are met: (a) content validity is adequate, (b) comparing individual scores are meaningful, (c) goals are scaled to "approximate interval scoring" (p. 245) and (d) the GAS quality criteria such as independent assessment of outcomes are met.

In the first assumption of the GAS, Kiresuk et al. (1994) describe that the content on the test must be adequate and representative of the content domain the tool intended to measure. They explain that the content on the test can only be judged if goals reflect the setting and the problems that may be relevant to the client in that setting. They claim that if the goals do not reflect client problems, the GAS does not show validity evidence pertaining to test content. Examination of test content however, typically needs content to evaluate, and the GAS is a measure with no fixed content. As each user of the GAS will outline different goals, the variability in goals makes it difficult to evaluate the content, since goals by themselves represent varying constructs. With the GAS, goals act as representations of the constructs that the treatment will address and they are indicators of the construct to be measured. Thus, evaluation of the GAS cannot proceed in the exact manner as measures that are content-laden or typical self-report questionnaires, and the lack of content in the GAS must be taken into consideration. These issues are ones that Kiresuk et al. (1994) acknowledge and in discussing the content limitation, they reinforce Messick's (1975) view that the use of test scores needs validation. More specifically, they draw the conclusion that the use of test scores need validation *not* the test content. This point is noteworthy and an important consideration; however, in order to evaluate the use of test scores, one also needs to examine the assumptions and rationales behind that score. Although the lack of fixed content has been noted by developers of the GAS, it is unclear how this aspect has been considered when validity evidence has been gathered about how the tool functions in practice and what justifications are provided for its use. Thus, a key consideration in gathering validity evidence are the processes that have been used to investigate validity for the GAS. In Chapter 2, I explore validation practices in a systematic review and collect information about how validity evidence for the GAS has been gathered since 1970.

Since the GAS has no fixed content or no preset "universe of content" (p.282), evaluating construct validity becomes imperative to adequately define what one is trying to measure (Cronbach & Meehl, 1955). Loevinger (1957) emphasizes content alone is not sufficient to illustrate whether a test measures a particular construct as she discusses the notion of substantive evidence. As noted by the *Standards*, studies of response processes are not limited to one respondent, as tests often rely on judges to evaluate one's performance (AERA et al., 2014). In the case of the GAS, the users typically include more than one individual to develop the content. In order to consider the joint processes that occur during administration and completion of the GAS, the nature of dyadic actions needs to be explored. As such, this dissertation uses APM from the position that goal setting and scaling is a joint project between a therapist and client is probable but not understood or verified. In Chapter 3, APM is used to explore what response processes emerge as dyads complete the GAS.

Altogether, by using the unified view of validity and an explanation-focused approach to validation, this dissertation examines how validity information has been gathered and what that information tells us about gaps in validation for the GAS (i.e. Chapter 2). The dissertation then introduces a new way to explore validity evidence based on response processes using APM (i.e. Chapter 3). By investigating substantive evidence, one is able to look closer at how the goal construct is represented in the measure, through process and content representation (Loevinger, 1957; Messick, 1995). Finally, in Chapter 4, some concluding thoughts are presented by summarizing findings, bringing concepts together, describing novel contributions, limitations and future implications.

Chapter 2. A systematic review of validation practices for the Goal Attainment Scaling measure[4]

2.1 Introduction

Goal attainment scaling (GAS) (Kiresuk & Sherman, 1968) is an internationally recognized measure that is used across many disciplines to identify and evaluate relevant goals for an individual (Kiresuk, Smith, & Cardillo, 1994). Although GAS originates from counselling and clinical settings, it is increasingly being used in educational settings, as goals and goal-setting are highly relevant to educational contexts and educational assessment (Kiresuk et al., 1994). The GAS is unlike typical measures since it lacks fixed content; that is, it consists of goals formed by the respondent and/or users in the process of completing the measure. The nonstandard format of the GAS has not deterred investigations of its measurement properties, such as validity and reliability evidence. This systematic review aims to understand validation practices for the GAS; specifically, how validity evidence is gathered and reported.

The GAS is often endorsed as an "individualized" measure since it allows users to develop and set personalized goals. At the time the GAS was first developed, it was used to specify individual goals for patients in a community mental health program. During this initial use, selection of goals for the GAS involved a committee or "goal selector" (p. 446) who determined a set of realistic goals for the patient, and then graded and scaled goals according to likely treatment outcomes (Kiresuk & Sherman, 1968). Later descriptions of the GAS modified this condition so goals were set either individually or collaboratively between a student and teacher, or client and practitioner (Kiresuk et al., 1994). Once goals are set, they are scaled to

[4] *Manuscript accepted at Journal of Psychoeducational Assessment*

identify variations in goal attainment, typically by individuals who have knowledge of the treatment or intervention (e.g. a practitioner or teacher). Scaling the goal involves identifying variations in goal attainment that indicates movement above or below treatment or intervention expectations. The GAS measure assumes that users will bring relevant or prior knowledge of the treatment or intervention to determine what goals are realistic, and to grade and scale these goals appropriately. Therefore, using the GAS involves: (a) assessing an individual's (e.g. student or patient's) skill level in a particular problem area, (b) developing and scaling a goal that is the intended result of a treatment or intervention, and (c) later scoring the goal based on perceived change. Altogether, the GAS is a unique measure and a striking feature is that it has "no fixed content," (Kiresuk et al., 1994, p.167), as users of the GAS determine both the goals and their scaling. Given this measure lacks fixed content and has been used for numerous years, examining validation practices will provide insight into how validity information is gathered for this unique measure.

Validity is defined as the justifications or explanations for variations in scores on a measure, and validation is the process of acquiring that information (Zumbo, 2009). Validity information provides evidence related to the content of tools that are used to measure phenomenon, as well as the interpretations and inferences that are made from their scores. While validity provides critical information about measures, it has also been reported that validation practices are inconsistent, and that there is a disconnect between the practice of validation and validity theory (Shankar, Miller, Roberson, & Hubley, 2019; Zumbo & Chan, 2014). Previous evidence has noted an imbalance in validity evidence presented and a *lack* of explicit reference to a validity framework (Shear & Zumbo, 2014). As the meaning and language surrounding validity has changed over the years, this may also influence how validity evidence for the GAS is

collected and reported since the original publication over 50 years ago (Kiresuk & Sherman, 1968). Common approaches for talking about validity include more than one view – modern (unified) validity theory and traditional (Trinitarian) validity theory (Guion, 1980; Newton & Shaw, 2013). The unified perspective was originally described by Cronbach and Meehl in 1955, and has evolved into a view that is currently endorsed by the *Standards* (AERA et al., 2014). This view of validity includes several sources of evidence and the *Standards* identifies five sources, which are: test content, response processes, internal structure, relations to other variables and consequences (AERA et al., 2014). Within the unified view, validity came to be seen as centered around the construct, with the sources of evidence all contributing to the "whole of validity" evidence (Loevinger, 1957, p. 636), and the importance of building a nomological network for interpretations of scores (Cronbach & Meehl, 1955; Hubley & Zumbo, 2011). The notion of the construct lies at the core of this view, whereby the term construct describes an unobserved concept or behavior that can be operationalized through a measurement process. In contrast, the assumption in the Trinitarian view is that validity exists as different "types" and this view sees validity as a property of a measure, so measures either do or do not have validity (Hubley & Zumbo, 1996). The tripartite view of validity has evolved and developed towards a more comprehensive view that considers validity as an integrative evaluative judgment, with validation as an ongoing process (Hubley & Zumbo, 2011; Messick, 1995). The unified view considers different types of validity (as historically considered), such as content and criterion-related, as subsumed under construct validity (Messick, 1989b). This review uses the unified view of validity as a guiding framework for studying validation practices by recognizing that all validity evidence contributes towards an understanding of the construct. The unified view is also

recognized by developers of the GAS (Kiresuk et al., 1994), who draw attention to the unified view of validity and discuss the importance of construct validation.

Of the sources of validity evidence identified in the *Standards*, test content and response processes are, arguably, foundational to the initial development and verification of a measurement instrument (AERA et al., 2014). Content related evidence evaluates how well the content in the instrument represents the construct it is intending to measure (AERA et al., 2014; Haynes, Richard, & Kubany, 1995; Sireci, 1998); and one can think broadly of response processes as "the mechanisms that underlie what people do, think, or feel when interacting with, and responding to, the item or task and are responsible for generating observed test score variation" (Zumbo & Hubley, 2017, p. 2), which in the case of the GAS, is connected to a goal set by the user. Evidence based on test content and response processes are complementary in their objectives and in their descriptions (Padilla & Benítez, 2014). They both evaluate the *representativeness* of a measure and its elements in relation to the construct by evaluating response consistency (Messick, 1975; Vogt, King, & King, 2004). Together these elements contribute towards an understanding of the meaning behind the GAS score (Messick, 1989a), and in particular what aspect of the goal construct the GAS intends to measure and how the GAS is interpreted among users.

Although the GAS is a measure that has variable content, some researchers have argued that evidence based on test content is a prerequisite for establishing other validity evidence (Vogt et al., 2004). Content related evidence can be defined as how a test is related to the content it is intended to measure, as well as the degree to which a measure represents a specific construct for a certain assessment purpose (AERA et al., 2014; Haynes et al., 1995). When evidence based on test content is obtained, the content domain for a measure is evaluated and feedback is received –

and it is through this process which justifies the content on the test, thereby judging the overall *quality* of a test (Sireci, 1998). Typically, test content evidence applies to the development and revision of instrument items, and the process includes specification of the construct of interest, review of test content and consultation with experts (Haynes et al., 1995; Vogt et al., 2004). Since the GAS does not have a defined "universe of items" (Cronbach & Meehl, 1955, p. 282) and instead relies on a universe of goals set by users, evaluating construct validity and specifying the construct of interest becomes essential to understand its content and adequately define what one is trying to measure. It has been suggested by Kiresuk et al. (1994) that various types of evaluation are limited by the GAS format. They contend that the final score in the GAS provides information about validity, which can be directly interpreted to evaluate change given certain assumptions are met. One of these assumptions is that content related evidence is "adequate" (Kiresuk et al., 1994, p. 245). As the content in the GAS is formed by users, evaluating test content must focus on the construct the GAS intends to measure to provide support and justification for subsequent score interpretations (Kane, 2006). Examining both the definition and operationalization of the construct are activities that gather content evidence in support of a measure's construct validity (Sireci, 1998; Vogt et al., 2004). Therefore, understanding what goal construct is being identified will highlight how the construct in the GAS is being operationalized and provide validity evidence towards the meaning of its score (Anastasi, 1986; Messick, 1980; Tenopyr, 1977).

Correspondingly, response process evidence examines the congruence between the construct and individual processes through examination of both theory and empirical analyses (AERA et al., 2014; Messick, 1989a). Response processes have traditionally been investigated using cognitive processing methods, such as think-aloud or cognitive interviews (Padilla &

Leighton, 2017). Examination of response processes have expanded to include aspects such as one's behaviour and motivations, to more fully understand what one is thinking or feeling as they interact with a measure (Leighton, Tang, & Guo, 2017). As well, aspects of emotion can be examined through various expressions (Leighton et al., 2017). In all, evidence based on response processes systematically assesses how respondents understand and process aspects of the construct that is measured by the GAS and can draw connections between the construct and individual responses. Along with evidence based on test content, these sources of validity evidence can provide justification for the goal construct the GAS purports to measure.

Synthesizing validity evidence provides a coherent account of the evidence supporting or disconfirming the intended interpretations from scores (O'Leary, Hattie, & Griffin, 2017). To appraise the way in which validity evidence is assembled, this review focuses on test content and response processes to examine representation of the goal construct measured by the GAS and to move beyond the individual 'test-takers' behaviours that are traditionally used in validation research, towards an explanation-focused view of validity (Zumbo, 2017a, 2017b). By applying this position, this systematic review examines validation practices for the GAS by collecting information about how validity evidence has been reported over the period 1970 to 2018. Specifically, the purpose of this systematic review is to investigate how validity evidence for the GAS is assembled and then examine the available validity and reliability evidence.

2.2 Methods

2.2.1 Search strategy

The following six databases were searched for relevant research articles: PubMed, Embase, Cumulative Index to Nursing and Allied Health Literature (CINAHL), Eric, PsycINFO, and Cochrane Database of Systematic Reviews. Library databases were examined using the

search criteria: (1) keyword "goal attainment scaling" combined with (2) valid* (* denotes truncation to search for variations in the word). Peer-reviewed articles, written in English, published since January 1970 that describe the use of GAS with any human sample were selected. Articles over several decades were searched to gather all available literature on validation practices with the GAS. Reference lists of articles that met full-text inclusion criteria were reviewed to determine if any additional articles should be retrieved.

2.2.2 Eligibility criteria

For inclusion in this review, articles were reviewed in a step-wise manner by two reviewers, once duplicates were removed: (1) titles and abstracts were screened, followed by (2) a review of the full-text to code and select final articles. Titles and abstracts were initially screened by the first author (SS) and a second reviewer (SKM) to determine whether articles identified: (a) use of the GAS as a measurement tool and (b) abstract identified measurement properties of the GAS. As valid* was already searched, if any measurement properties were mentioned, the terms validity or validation would be contained in the full text. Altogether, this process was liberal and tended towards inclusion rather than exclusion. The focus was to include all articles and examine how validity was conceptualized and examined. Furthermore, experimental studies, reviews and commentaries were all included to examine how validity evidence for the GAS was both investigated as well as interpreted. As a final screen, the first author (SS) reviewed the full-text before selecting final articles; in this last step, all articles were coded to examine how validation evidence was investigated or described in each corresponding article. The second reviewer (SKM) reviewed 20% of full-text articles to verify the coding and the article selection process, and to obtain a macro-level sense of similarities of ratings between both raters. Final articles were selected if the coding process verified validity evidence was

examined or reviewed with the GAS. Only studies that examined validity as it pertained to measurement properties of the GAS were included, and articles that discussed: social validity, ecological validity or treatment validity were excluded. Furthermore, articles that only mentioned the GAS as "valid" but did not describe specifics about the validation process were excluded. Once data were read and coded, only those articles that described measuring validity or providing validity information for the GAS instrument were included. Figure 2.1 outlines details about article selection and reasons for exclusion in the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) protocol (Moher, Liberati, Tetzlaff, Altman, & Group, 2009).

2.2.3 Coding

A data abstraction and coding form was developed by authors to gather validity and validation information from screened articles. A coding sheet that included two sections was used to assess the available validity evidence. Section A collected descriptive information about articles, sample characteristics and noted which articles identified as a review by their synthesis of literature, and Section B collected information about validity and validation evidence.

Using a similar coding process as Chan et al. (2014) and Cizek, Rosenberg, and Koons (2008), Section B was organized using the following scheme and the five sources identified in the *Standards* (AERA et al., 2014) (a) test content, (b) relations with other variables, (c) internal structure, (d) consequences, and (e) response processes (see Appendix C for a sample of the coding sheet and definitions). As well, an "other validity" category was included to account for validity sources not described in the *Standards* (AERA et al., 2014) and reliability was also included, since reliability is a necessary condition for validity (Hubley & Zumbo, 2013). In particular, the type of reliability estimate, specifically internal consistency, test-retest or inter-

rater estimates were noted to make a comparison with the amount of validity evidence that was sought and reported. Furthermore, to understand validity perspectives this review coded for the following: (a) whether articles mentioned a unitary perspective of validity and (b) whether articles stated that validity was a property of the test or a property of test scores.

2.2.4 Data analysis

Validity evidence that was gathered from coded articles was evaluated using a combination of inductive and deductive content analytic approaches (Elo & Kyngäs, 2008; Hsieh & Shannon, 2005; Mayring, 2000). Gathering descriptive information invoked a deductive approach as information reported in research articles were noted. A deductive content analytic approach was used to collect information on all sources of validity evidence. For instance, when gathering construct definitions for evidence based on test content, as well as evaluating the use of theory and systematic testing of individual processes provided evidence for response processes. A deductive content analytic approach was also used when information regarding validation practices was applied to coding. Evaluating sources of validity evidence and validity perspectives included a combination of approaches; a deductive approach was used to collect information that was reported and an inductive approach was used to interpret this information.

2.2.5 Inter-rater agreement on coded studies

The first author and second reviewer were in agreement for 50 of 60 data points for these 37 studies, and this represented 83.3% agreement. Any differences were discussed between the first and second rater to reach consensus before assigning codes to cases.

2.3 Results

This review identified a total of 115 articles once duplicates were removed. After abstracts were screened and full text reviewed, a total of 37 articles were selected as examining

or reporting validity evidence for the GAS (Figure 2.1). Of the selected articles, 10 identified themselves as a review and synthesized literature. Selected articles examined the GAS for a variety of reasons related to validity or validation, such as: to gather information about validity and/or other measurement properties, examine the feasibility or utility of the GAS in certain settings, use GAS as an outcome measure and review the GAS by itself or in comparison to other measures. The GAS was used with individuals who were patients or students. The samples included: children, youth, adults and older adults, and in one article the sample was nonspecific (Table 2.1).

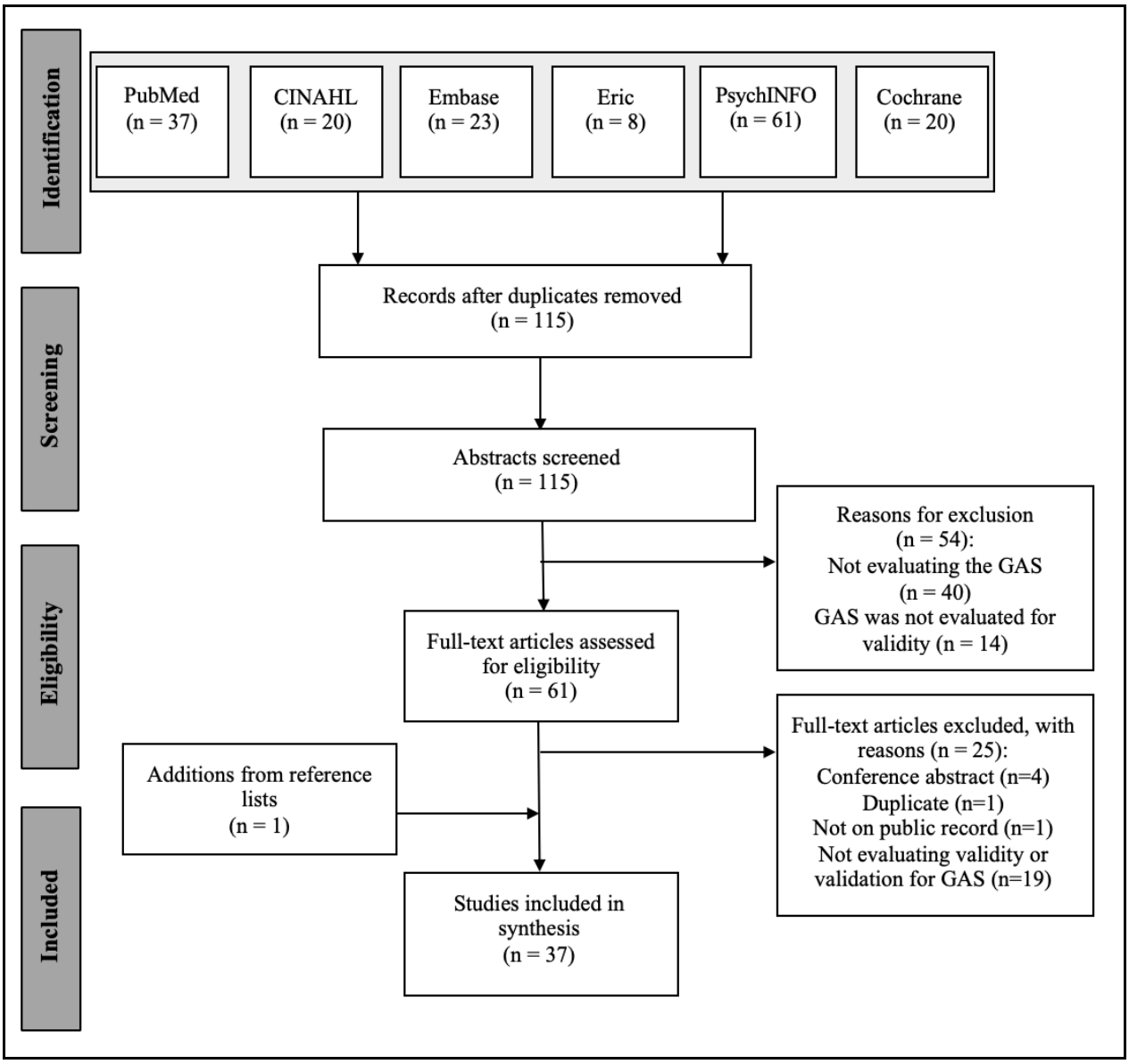

Figure 2.1 PRISMA flowchart for systematic review

Table 2.1 Article Description

Author	Review article?	Article Title	Sample
Bouwens, van Heugten, & Verhey (2009)	Yes	Review of Goal attainment scaling as a useful outcome measure in psychogeriatric patients with cognitive disorders.	Older Adults
Calsyn & Davidson (1978)	No	Do we really want a program evaluation strategy based solely on individualized goals? A critique of goal attainment scaling.	Adults
Cusick, McIntyre, Novak, Lannin, & Lowe (2006)	No	A comparison of goal attainment scaling and the Canadian occupational performance measure for paediatric rehabilitation research.	Children
Cytrynbaum, Ginath, Birdwell, & Brandt (1979)	Yes	Goal attainment scaling: A critical review.	Adult and Children
de Beurs et al. (1993)	No	Goal attainment scaling: An idiosyncratic method to assess treatment effectiveness in agoraphobia.	Adults
Donnelly & Carswell (2002)	Yes	Individualized outcome measures: A review of the literature.	Adults
Fisher & Hardie (2002)	No	Goal attainment scaling in evaluating a multidisciplinary pain management programme.	Adults
Gaasterland, Jansen-van der Weide, Weinreich, & van der Lee (2016)	Yes	A systematic review to investigate the measurement properties of goal attainment scaling, towards use in drug trials.	Various
Gordon, Powell, & Rockwood (1999)	No	Goal attainment scaling as a measure of clinically important change in nursing-home patients.	Older Adults
Heavlin, Lee-Merrow, & Lewis (1982)	No	The psychometric foundations of goal attainment scaling.	Nonspecific
Hurn, Kneebone, & Cropley (2006)	Yes	Goal setting as an outcome measure: A systematic review.	Adults & Older Adults
Jones et al. (2006)	No	Using goal attainment scaling to evaluate a needs-led exercise programme for people with severe and profound intellectual disabilities.	Adults
Joyce, Rockwood, & Mate-Kole	No	Use of goal attainment scaling in brain injury in a rehabilitation	Adults

Author	Review article?	Article Title	Sample
(1994)		hospital.	
Kiresuk, Lund, & Larsen (1982)	No	Measurement of goal attainment in clinical and health care programs.	Adults and Children
Krasny-Pacini, Evans, Sohlberg, & Chevignard (2016)	No	Proposed criteria for appraising goal attainment scales used as outcome measures in rehabilitation research.	Adults and Children
Krasny-Pacini, Hiebel, Pauly, Godon, & Chevignard, (2013)	Yes	Goal attainment scaling in rehabilitation: A literature-based update.	Adults and Children
Malec (1999)	No	Goal attainment scaling in rehabilitation.	Adults
Mannion, Caporaso, Pulkovski, & Sprott (2010)	No	Goal attainment scaling as a measure of treatment success after physiotherapy for chronic low back pain.	Adults
Mcgaghie & Menges (1975)	No	Assessing self-directed learning.	Students
Palisano & Gowland (1993)	No	Validity of goal attainment scaling in infants with motor delays.	Adults
Palisano, Haley, & Brown (1992)	No	Goal attainment scaling as a measure of change in infants with motor delays.	Children
Rock (1987)	No	Goal and outcome in social work practice.	Adults and Children
Rockwood (1994)	Yes	Setting goals in geriatric rehabilitation and measuring their attainment.	Older Adults
Rockwood, Stolee, Howard, & Mallery (1996)	No	Use of goal attainment scaling to measure treatment effects in an anti-dementia drug trial.	Adults
Rushton & Miller (2002)	No	Goal attainment scaling in the rehabilitation of patients with lower-extremity amputations: A pilot study.	Adults
Sakzewski et al. (2007)	Yes	Clinimetric properties of participation measures for 5- to 13-Year-Old children with cerebral palsy: A systematic review.	Children
Schlosser (2004)	No	Goal attainment scaling as a clinical measurement technique in communication disorders: A critical review.	Adults and Children

Author	Review article?	Article Title	Sample
Shefler, Canetti, & Wiseman (2001)	No	Psychometric properties of goal-attainment scaling in the assessment of Mann's time-limited psychotherapy.	Adults
Steenbeek, Ketelaar, Galama, & Gorter (2007)	Yes	Goal attainment scaling in paediatric rehabilitation: A critical review of the literature.	Children
Stolee et al (2012)	No	The use of goal attainment scaling in a geriatric care setting.	Older Adults
Stolee, Rockwood, Fox, & Streine (1992)	No	A multi-site study of the feasibility and clinical utility of Goal Attainment Scaling in geriatric day hospitals.	Older Adults
Stolee, Stadnyk, Myers, & Rockwood (1999)	No	An individualized approach to outcome measurement in geriatric rehabilitation.	Older Adults
Turner-Stokes, Fheodoroff, Jacinto, Maisonobe, & Zakine (2013)	No	Upper limb international spasticity study: Rationale and protocol for a large, international, multicentre prospective cohort study investigating management and goal attainment following treatment with botulinum toxin A in real-life clinical practice.	Adults
Vu & Law (2012)	Yes	Goal-attainment scaling: A review and applications to pharmacy practice.	Adults
Willer & Miller (1976)	No	On the validity of goal attainment scaling as on outcome measure in medical health.	Adults
Woodward, Santa-Barbara, Levin, & Epstein (1978)	No	The role of goal attainment scaling in evaluating family therapy outcome.	Adults
Yip et al. (1998)	No	A standardized menu for goal attainment scaling in the care of frail elders.	Older Adults

2.3.1 Reporting of validity evidence

The majority of articles in this review reported on or examined relations to other variables (89.2%), followed by evidence based on test content (51.4%) (Table 2.2). Evidence based on consequences was reported in one article (2.7%) and no articles reported information about internal structure or response processes.

Table 2.2 Frequencies of Sources of Validity Evidence Reported

Validity Evidence	# of articles (N)	Percentage (%)
Test content	19	51.4
Response processes	0	0
Internal structure	0	0
Relations to other variables	33	89.2
Consequences	1	2.7

2.3.1.1 Construct representation through test content and response processes

To gather information about evidence based on test content, articles were reviewed to see if experts were consulted, a construct was identified, and if a corresponding construct definition was provided. Content was evaluated by examining goal domains, agreement between experts, comparing goals on the GAS to the content from other reports or assessments, and expert opinion. Expert panels included practitioners, patients, students, family, team members or individuals doing intake. Although evidence based on test content included some expert consultation, the construct measured by the GAS was not clearly identified. In several articles, there was no clear construct that the GAS was identified as measuring, and articles identified the

GAS in two predominant ways (Table 2.3): (1) solely as a measure of goals (40.5%), or (2) as both a measure of goals and its own method or measurement technique (56.7%), and one article stated the GAS construct lacked clarity and was nonspecific. While several articles mentioned a theory (13.5%) and one article identified a specific goal theory, no articles actually used a theoretic approach.

To examine evidence based on response processes, this review evaluated whether theory was used to guide application of the GAS, and also whether response processes were empirically tested. Among the articles that mentioned a theory, only one article (i.e. Hurn, Kneebone, & Cropley, 2006) identified a specific goal theory. Furthermore, results indicate that no articles reported information about systematic testing of response processes, such as cognitive, motivational or behavioural types of processing. Nonetheless, many articles mentioned observations or reflections upon how goals were set, such as through negotiation or consensus.

Table 2.3 Description of the GAS

GAS description	**# of articles (N)**	**Percentage (%)**
Measure of goals (includes: goal achievement/attainment or goal-setting)	15	40.5
Measure of goals *and* its own measurement technique or approach	21	56.8
Nonspecific	1	2.7

2.3.1.2 Internal structure, relations with other variables, and consequences

No articles reported evidence based on internal structure of the GAS which is congruent with this type of measure since it lacks fixed content. Validity evidence based on relations with other variables was reported in varying ways, such as construct, convergent, concurrent, criterion, and predictive validity, as well as the nomological network. In addition, responsiveness or sensitivity to change was reported in over half of the articles (67.6%). As well, the only article (i.e. Rockwood, 1994) that reported considering the unintended consequences of testing, advocated for the individualized nature of the GAS as guarding patients against unintended consequences of other measures. Among reviewed studies that mentioned a score and its applied purpose, almost half of all articles (48.7%) discussed that the GAS score may be interpreted as a change score.

2.3.2 Other evidence related to validity

Numerous articles (37.8%) also identified other sources of validity evidence that were outside the criteria identified, such as face validity, external validity, internal validity and congruent validity.

2.3.2.1 Reliability

Although all included articles discussed validity evidence of the GAS, the majority of articles also reported or examined corresponding reliability evidence (94.6%). Inter-rater reliability was reported most frequently (73.0%), followed by test-retest (16.2%) and internal consistency (13.5%).

2.3.3 Validation practices

Based on the validity evidence that was gathered, most studies tended to gather validity evidence by focusing on relations with other variables and reliability evidence. Validity evidence was often gathered as types and studies tended to gather different types of validity. No articles mentioned the unitary perspective of validity or any editions of the *Standards* (e.g. AERA et al., 2014), and one article identified the tripartite view (i.e. Gordon, Powell, & Rockwood, 1999) as their theoretical approach to validity. Validity was discussed as either a property of the GAS measure or the GAS scores.

2.4 Discussion

This review provides a glimpse into validation practices for the GAS measure by examining how this evidence was gathered and assessing available validity evidence. The 37 articles selected for this study verified that the GAS is used in a variety of settings and with a variety of samples (Table 2.1).

2.4.1 Validity and validation evidence of the GAS

Validity evidence was evaluated by all studies in a number of ways. Most validity evidence tended to focus on 'relations to other variables,' or reliability. The concentration of data in these areas is not uncommon and is similar to the findings by Zumbo and Chan (2014). They note that the high concentration of this evidence brings some difficulty interpreting evidence and building a sound validity argument. They also note limited guidance from orientations to theory, including validity theory. Altogether, validity evidence was not gathered in any systematic way and reflected a piecemeal approach to validation.

Articles approached validity by gathering varying *types* of validity evidence and then reasoning that the GAS was either 'valid' or 'not valid.' Discussing validity in this way, as a

property of the measure or its scores, implies that validity is seen as a fixed or immutable quality. This idea emerged in the 1940s, where validity was conceptualized as a static property that had to be proven or established (Goodwin & Leech, 2003). Newton and Shaw (2013) in their analysis of ways validity is discussed or thought about, suggested validity was referred to as if it were a property of a test for several reasons, such as: intentional misuse, lack of awareness or misunderstanding, and genuine divergence from the view of validity as a property of interpretations. Newton and Shaw (2013) also identified 122 discrete validity labels intended to capture an aspect of validity for measurement. From these labels and the results from this review, it is apparent that articles do not consider validity as the *interpretations from scores* on a measure. As discovered in this review, the GAS has varying interpretations.

2.4.2 Interpretations of the GAS

The GAS was frequently identified as both a measure of goals and also its own measurement technique. Among articles that described the GAS solely as a measure of goals, articles either specified goals were individualized, or simply referred to goals in a broad or general manner. The term *goal* was used by itself or referred to related aspects of the goal construct, such as goal achievement, goal attainment or goal-setting. It was difficult to decipher how varying aspects of the goal construct were distinguished since these terms were used interchangeably in relation to the GAS.

2.4.2.1 Goal constructs identified in the GAS

The results from this review draw attention to the various ways goals may be described in the GAS, and some discrepancies between how aspects of the goal construct are discussed. One discrepancy is that individualized goals or goal achievement is not equivalent to the process of goal-setting. However, studies included in this review readily moved from identifying the GAS

as measuring goals, to the GAS as a tool for goal-setting and also evaluating goal achievement. Ostensibly, these aspects or dimensions of the goal construct were all viewed somewhat synonymously.

Although similarities exist between these various aspects of the goal construct, there are important distinctions that influence what outcomes are produced from the GAS, as well as the score meaning. Of the articles included in this review, only one article (e.g. Hurn et al., 2006) identified a construct definition for a goal. As noted by Elliot and Fryer (2008), the term goal is rarely defined in research, with the assumption that all readers understand the word similarly, but the term can take on different meanings. While it is not uncommon to consider the goal construct in a parsimonious way, separating its differential aspects has a number of advantages, such as: limiting the confusion surrounding the construct, improving understanding of the influence of multiple goals, and minimizing assumptions (Austin & Vancouver, 1996).

As noted by Austin and Vancouver (1996), proliferation of various aspects of the goal construct makes its examination problematic, which is evidenced in this review. Presumably, the GAS has been assumed to measure different aspects of the goal construct and no studies examining validity evidence considered their distinctions. Improving the clarity surrounding the goal construct in the GAS can help determine clear functional properties of this measure, instead of wondering if additional constructs may account for a particular behaviour (Elliot & Fryer, 2008). Likewise, a number of factors affect goal outcomes, their structure and process (Austin & Vancouver, 1996). For instance, the activation and pursuit of goals depends on one's conscious desires, which can influence their thoughts, emotions and behaviors (Fishbach & Ferguson, 2007); and goal commitment is only recognized when there is an investment of affect, cognitive resources and behavior (Mann, de Ridder, & Fujita, 2013). Moreover, goal achievement is

influenced by the nature of the task and how applicable the goal is towards it (Fishbach & Ferguson, 2007), as well as the context and one's level of control (Austin & Vancouver, 1996). With a vast amount of psychological literature surrounding the goal construct, these aspects reflect a succinct view into some of the factors associated with this construct. Certainly, from a validity and validation standpoint, the varying constructs identified suggest clarity is needed to enhance explanations and score meaning.

2.4.2.2 Evidence of theory guiding definitions

Validity is a matter of inference and a process that provides information related to the meaning of scores, which in turn provides information about an outcome of interest. Understanding what inferences are made with test scores refers back to how theory is applied to justify the claims that are made regarding these scores (Kane, 2006). The only critique included in this review wondered, "*How is the construct embedded in the theory?*" (Cytrynbaum, Ginath, Birdwell, & Brandt, 1979, p.33) and the results from this review indicate that this question remains unexamined and therefore unverified. Theoretical perspectives about the goal construct are rarely mentioned and notably absent in applications of the GAS. Among studies reviewed, only one article (e.g. Hurn et al., 2006) specifically mentioned a goal theory (i.e. Locke, 1968), but this was not operationalized in the respective study. Another study included in this review mentioned that theory relates to something clinicians consider during the goal-setting process, however, this too was not tested (e.g. Vu & Law, 2012). Using the GAS to set goals is a complex endeavour, and theoretical rationales can assist by providing guidance in action-planning this process (Scobbie, Dixon, & Wyke, 2011). In a Cochrane review that investigated the GAS and goal-setting in rehabilitation medicine, the authors noted that only one study implemented goal-setting in a way that was consistent with a theory (Levack et al., 2016). The results from this systematic

review suggest that future research using the GAS needs to better implement a theory to guide establishment of a definition and its application.

Given the GAS does not have items like a conventional measure that is scored and lacks fixed content, this added complexity stresses the importance of construct definitions and theory to guide how this construct is operationalized. By discounting how these aspects contribute to validity evidence, one can only be certain that they are *assuming* the GAS measures an aspect of the goal construct, not verifying it, which has consequences for users of the GAS. There is no shortage of applicable theories that relate to the goal construct (Austin & Vancouver, 1996) and theories of behavior change can inform goal-setting interventions (Scobbie et al., 2011), such as: Social Cognitive Theory (Bandura, 1997), Goal Setting Theory (Locke & Latham, 2002), and the Health Action Process Approach (Schwarzer, 1992), all of which could be applicable to the GAS. Indeed, from a validation perspective, strong forms of construct validity evidence include a theory that is well-articulated and tested, which helps to strengthen the nomological network and provide a sound validity argument (Cronbach & Meehl, 1955; Loevinger, 1957; Messick, 1989b; Zumbo, 2009).

2.4.2.3 The need for evidence based on response processes

In addition to theory, this systematic review also examined whether individual interactions with the GAS were tested, as an aspect of response process evidence. No articles tested response processes, including the more commonly examined cognitive processes, such as think aloud or cognitive interviews (Padilla & Leighton, 2017). Nonetheless, articles did mention aspects related to how individuals interacted with the measure and how goals were set. Articles included in this review often stated whether goals were set collaboratively or alone and how final goals were determined (e.g. through negotiation or consensus). This review uncovered that several

articles considered some aspects related to goal-setting, but no articles empirically tested and verified these aspects. For instance, two articles contemplated student or patient feelings and motives (e.g. Mcgaghie & Menges, 1975; Stolee et al., 2012), the patient or family concerns (e.g. Stolee, Stadnyk, Myers, & Rockwood, 1999), and in another case an article addressed conceptualizing goal-setting as difficult for a cognitively impaired patient (e.g. Krasny-Pacini, Evans, Sohlberg, & Chevignard, 2016). As well, one article stated that goal orientations can be influenced by an individual's motivation (e.g. Kiresuk, Lund, & Larsen, 1982), and another mentioned that precision of goals was related to reporting and how goals were identified (e.g. Milne, Robert, Tang, Drummond, & Ross, 2009). It is noteworthy that these aspects were considered; however, testing these considerations by investigating individual interactions with the GAS can provide empirical evidence to support these claims.

2.4.3 The GAS score and its meaning

Altogether, building a validity argument is a key aspect of strengthening score interpretations. Although reliability is a part of the validity argument and provides insight into the consistency of the GAS scores; it contributes minimally to the accuracy of the findings and is not a substitute for validity (Barry, Chaney, Piazza-Gardner, & Chavarria, 2014; Zumbo & Chan, 2014). Reliability was reported in almost all reviewed articles; however, it is not enough to justify the use of the GAS score. A fundamental feature of the validity argument and integrating validity evidence within a unitary concept of construct validity is how the construct is represented (Messick, 1995).

Almost all the studies included in this review mentioned the GAS score was measuring a change, and an applied purpose of the GAS score was to produce a change score. In most cases, change was measured with respect to student or patient progress and to compare program

effectiveness – e.g. "program success is measured in "goals achieved,"" (Calsyn & Davidson, 1978, p.306). In addition, it was not uncommon for studies to interpret a particular GAS score as an evaluation of change (i.e. improvement, no change or deterioration). Articles discussed a number of different ways the GAS evaluates change and used the terms: responsiveness, sensitivity to change and change score to discuss or denote change over time; reporting of change was highly variable and inconsistent. In a literature review that investigated how studies of treatment effectiveness and program evaluations measure change over time, the authors found there are challenges to interpreting change scores, and difficulty comparing estimates of responsiveness (Beaton, Bombardier, Katz, & Wright, 2001). Beaton et al. (2001) also noted that the same terms were used, sometimes interchangeably, and emphasize that statistics like responsiveness are highly contextualized. Responsiveness is a term that is widely used to denote change over time or sensitivity to change (Middel & van Sonderen, 2002; Terwee, Dekker, Wiersinga, Prummel, & Bossuyt, 2003) and refers to the ability of a measure to accurately detect change when it has occurred (Beaton et al., 2001). While change scores are also considered as an indicator of change over time (Thomas & Zumbo, 2012), they may also refer more generally to any difference score (Cronbach & Furby, 1970). Although the GAS produces a score that incorporates different time points, the nature of the change needs to be understood so interpretation of the GAS score as demonstrating change is reasonable and clear.

Thus, questions about whether the GAS is a suitable measure to evaluate change hinges on clarity about its construct definition and subsequent score meaning. Before one can determine whether the GAS is a measure of change and can effectively measure change in a goal construct, understanding the interaction between participants' responses and how they align with the goal construct is imperative. As noted by one article included in this review, "initial exposure to goal

setting may have allowed the person time to reflect, thereby possibly leading to a change in the goal areas" (Rushton & Miller, 2002, p.776). There are a number of reasons goals may change, as well as factors that influence their achievement. Although many factors relate to the goal construct, perhaps most important is a well-identified gap between goal intentions and goal behaviour, demonstrated in a meta-analysis by Webb and Sheeran (2006). This discrepancy suggests that intentions do not necessarily lead to actions, and actions need facilitation (Webb & Sheeran, 2006). Indeed, activation of a goal can dissipate once a goal has been reached or if an obstacle that cannot be overcome is encountered (Fishbach & Ferguson, 2007), and activation of multiple goals can shift over time (Austin & Vancouver, 1996). Thus, if the GAS is, as Kiresuk et al. (1994) maintain, a measure of one's "perceived ability to change" (p. 245), how does one know if the GAS is measuring change in the identified goal or quite simply a change in goal?

2.4.4 Strengthening validity evidence and validation practices

Ultimately, a score cannot be interpreted on a test if one does not know what the test is measuring (Sireci, 2012), and as shown through variations in goal construct, 'what the test is measuring' is unclear. Of the 37 articles included, 10 identified themselves as a review article of GAS literature; however, none appraised interpretations of the GAS or drew connections to theory. Strengthening the validity argument for the GAS requires better testing procedures in order to justify the goal construct measured by the GAS and verify that the GAS does measure change. As well, consequences of score interpretation and use was considered in one article (e.g. Rockwood, 1994) and more studies need to consider the applied purpose of the GAS score. Effectively, validation requires scientific inquiry alongside a rational argument to substantiate the score interpretation and use (Messick, 1995).

The variability noted among reviewed articles highlights that the GAS has many different interpretations. There is a lack of clarity regarding how the GAS is best interpreted, what specific construct the GAS measures and whether the GAS measures a goal construct or whether the GAS is best regarded as its own measurement technique. An advantage of the unified view is that score interpretations infer a construct that underlies their score (Sireci & Sukin, 2013), and this logic can improve how the GAS is used and discussed. Given the GAS does not have items like a conventional measure that is scored, this added complexity stresses the importance of construct definitions and theory to guide how this construct is operationalized. Importantly, researchers need to gather input from students, patients and/or families as part of response process information and validation efforts. Evidence based on response processes will enable researchers to link theoretic information and judgments about the content of a test with consistencies in item responses; thus improving explanations of score meaning and subsequent interpretations, as well as the consequences of testing (AERA et al., 2014; Messick, 1989a).

This review noted that validity evidence was often provided without explanations that enhance interpretations of the GAS measure. Articles did not employ the unitary perspective of validity that has been encouraged by the *Standards* (AERA et al., 2014), and did not regard validity as an integrative judgment (Messick, 1995; Zumbo, 2009). The way in which evidence was gathered and presented, suggest some crucial changes are necessary to update measurement knowledge across disciplines, for better implementation and stronger collaborative practice. Perhaps most important in outlining a sound validity argument is for researchers to begin by identifying a validity theory to guide their validation approach. Researchers may legimately choose to use one validity approach over another; however, in its absence, and as shown in this review, validity evidence does not move towards the same objective. At the core of our findings

is that the construct measured by the GAS is unclear and has not been substantiated by previous validity evidence. While differences will continue to exist between researchers and disciplines in choosing one view over the other, an obvious question to ask is '*how validity evidence enhances our understanding of the GAS?*' or any measure, for that matter.

2.5 Conclusions & Recommendations

This systematic review is a unique contribution to the interdisciplinary measurement literature and highlights some gaps in the accumulated validity evidence for the widely used GAS across disciplines. This investigation goes beyond studies that simply conducted examinations of validity; I synthesize validation practices and highlight gaps in evidence which limit confidence in the GAS. Fundamentally, the inability to identify a clear goal construct for the GAS impacts the ability to measure this construct reliably and suggests some core aspects that are problematic. This review demonstrates the importance of building a validity argument starting with identifying a validity approach, and points out the influence of theory and response processes to substantiate the construct in the GAS measure. Use of the *Standards* (AERA et al., 2014) is recommended as a decision-making tool to strengthen validation practices. Its use is encouraged to improve how validity evidence is considered and gathered, and should not be mistaken as a check-box list of guidelines one follows mechanically.

In addition to investigating validation practices and validity evidence for the GAS measure, this review shows that validity evidence for test content and response processes are key pieces of evidence in establishing what construct the measure represents. This review found that no articles questioned whether applying approaches to examine validity for measures with specific items should be applied to the GAS; a measure in which the content is formed by the respondent and/or users during completion of the GAS. Instead, articles applied the same

procedures as are commonly used for measures with fixed items to examine validity. Consequently, this review provides a never before seen look into measures without uniform content, and opens several opportunities for future validity research.

It is often emphasized that the GAS is a measure of change and the score indicates change in goal attainment. Therefore, the GAS score has an applied purpose and a social consequence (Messick, 1989a). Whether the GAS score is used in educational or clinical settings, its score has meaning and a judgment or interpretation is formed based on its value. In order to evaluate how plausible an interpretation is, "it is necessary to be clear about what the interpretation claims" (Kane, 1994). This review points to areas for further improvement in validity evidence for the GAS and urges researchers to consider ways validation practices can help verify the many claims that are made about this measure.

Chapter 3. Investigating response processes for the Goal Attainment Scaling measure using Action-Project Method

3.1. Introduction

Responding to items on a measure is a complex human endeavor and understanding how and why people respond the way they do occurs through examination of response processes. Examining the aspects that underlie what individuals are doing means understanding how individuals engage with the tool; it means considering aspects such as cognitive processes (e.g. thinking), emotion, motivations, affect, actions, and behaviour (Embretson, 2016; Hubley & Zumbo, 2017; Leighton & Gierl, 2007). Current methods of gathering response process data only capture cognitive processes and are unable to investigate processes when more than one user are engaged with a test. This article tests and gathers response processes data from occupational therapists by using a modified version of the Action-Project Method (APM; Marshall, Zaidman-Zait, Domene, & Young, 2012; Young, Valach, & Domene, 2005) to build validity evidence for a goal-setting measure, Goal Attainment Scaling (GAS; Kiresuk & Sherman, 1968) in the context of a therapeutic interview with a client. Starting with an introduction to the GAS, this article explains the need for response process data for this and other measures that involve joint processes, and then explains how APM can be used to investigate response processes.

The GAS was originally developed for use in mental health settings, but has also been used in education, social work and counselling settings (Kiresuk, Smith, & Cardillo, 1994). It is a widely used international tool in which users set, scale and score relevant goals. A notable feature of the GAS is that its content, which are goals, are formed by respondent(s) during use of the tool. Kiresuk et al. (1994) describe that goals can be set by a therapist or intake worker,

client, or both the therapist and client. Goal setting between dyad members typically occurs through unstructured interviews during which a therapist sets goals based on discussions with a client as well as prior knowledge, and then scales these goals by considering the results of the treatment or intervention plan. The goals that are set and scaled are expected to correspond to client performance as a result of treatment or intervention, and the resulting score provides interpretations about one's performance and "degree of change" (Kiresuk et al., 1994, p. 5). There is no set format or structure for how goals are set using the GAS as the process is open-ended. Thus, goals will be different among users and across settings. However, once goals are set, they are scaled and later scored during a follow-up interview to evaluate perceived change. According to the GAS instructions (cf. Kiresuk et al., 1994), when users are engaged with the GAS measure, they need to start by setting a goal. Then users set an expected level of outcome for the goal, followed by outcomes that highlight somewhat more and somewhat less than expected, and finally outcomes that are much more and much less than expected. These outcomes are scaled so that the expected level is 0, somewhat more or somewhat less than expected is +1 or -1, and much more or much less than expected is +2 and -2, respectively. A fundamental assumption associated with this tool is that it measures a client's goals resulting from treatment. In particular, it is assumed that *goals* are being set and scaled when users, such as a therapist, interact with the GAS, but *how* this process occurs has never been examined or verified. Understanding why an individual responds in a particular way and why it is scaled in a certain way would help bridge the gap between scores on the measure and the inferences that are made about the score (Zumbo, 2009).

Although a number of studies have examined validity evidence for the GAS, a recent review by Shankar, Marshall, and Zumbo (*in press*) found validation practices vary considerably

and no studies have examined response processes. This systematic review also found that goals were set in a variety of ways and theory was not applied to guide use of the GAS. Altogether, the results suggest limitations in validity evidence and the inferences that can be made from the GAS (Shankar et al., *in press*). Even though it remains unclear how individuals use the GAS, engagement with the GAS represents an interactive situation between the measure, as well as the specific context. As users are engaged with the GAS, identification of response processes indicate how they are engaging with the measure and what response processes are involved. Specifically, response processes refer to an analysis of individual responses (AERA et al., 2014) that, "go beyond the surface content of the actions, thoughts, or emotions" (Zumbo & Hubley, 2017, p.3). They move beyond observed or expressed responses and include aspects such as cognition or motivation. During use of the GAS, response processes can indicate whether engagement with the GAS is solely goal-specific or related to other factors. For instance, during engagement with the GAS, therapist-users may focus on goals relevant to treatment or may be more involved with interacting with the client. Furthermore, users may also be involved in various cognitive or psychological processes to integrate information to set and scale relevant goals.

According to instructions provided by authors of the GAS (Kiresuk et al., 1994), users must proceed through a number of steps in order to set and scale goals. At this time, it is unknown how therapists and clients actually set goals and whether these goals are set jointly. The main user, typically a therapist must direct the process and start by identifying issues that will be addressed in treatment, translating the problems into 3 goals and specifying outcome levels that correspond to expected goal achievement for a client (steps 1-9 in Kiresuk et al., 1994, p. 7). In order to operationalize these steps, the therapist must do a number of things

concurrently, such as: (a) utilize prior knowledge based on their training, (b) consider the treatment and progression of treatment or intervention, (c) engage with the client, (d) predict expected level change for a client and (e) take into account other relevant information. Altogether, examining the ways through which a therapist begins to operationalize instructions of this measure will provide information about the nature of dyadic goal construction using the GAS. Thus, understanding the response processes involved during application of the GAS will provide evidence towards how respondents are engaged with this tool. Using this information, the current study attempts to gather response process data for the GAS measure.

Gathering response process data occurs through a number of methods. Typically, these methods have focused on cognitive processes, such as think aloud protocols, cognitive interviewing and Cognitive Aspects of Survey Methodology (Tourangeau, 1984). Other methods also include eye tracking or analyzing how components of a test or task are related. The aforementioned methods are typically used with surveys, tests or measures that have structured components or indicators, such as specific questions or items. However, during the use of the GAS in dyadic interactions, a therapist engages in a number of concurrent processes in an unstructured format. Thus, evaluating a tool where content is formed by user(s) requires a method that offers insight into response processes while setting and scaling goals.

One approach that is suitable for dyadic or triadic contexts and offers a way to examine response processes and the nature of dyadic interaction with the GAS is the Action-Project Method (APM; Marshall et al., 2012; Young et al., 2005). Applying APM may uncover the series of actions that occur between a dyad in completing the GAS and illustrate various perspectives of action for this joint project (Young et al., 2005). It is expected that using APM will enable a nuanced glance into joint projects, whereby goal setting and scaling can be

examined in the context of a conversation between a therapist and a client. APM is conceptually grounded in action theory, which understands action as inherently goal-directed and purposive although not necessarily rational (von Cranach et al., 1982). In particular, APM draws on contextual action theory (Young, Valach, & Collin, 2002) which provides a framework to support relational processes and joint actions (Young et al., 2001). Action theory considers action from three perspectives (Domene et al., 2015): (a) manifest behaviour – which is the observable behaviour necessary to carry out an action, (b) internal processes – the cognitive and emotional processes experienced during an activity and (c) social meaning – which are the explanations about one's actions that also consider the intentions and purpose of the action. The perspectives considered by action theory, such as manifest behaviour and internal processes align with aspects examined by response processes. As action theory considers action as inseparable from, and integrative of cognitive processes, emotion, psychological processes, social meaning and one's intentional stance (Valach et al., 2015), APM offers a method to examine response processes. Furthermore, an advantage of action theory is its applicability towards actions that are intentional and nonlinear, as well as its consideration of dyadic joint actions (Marshall et al., 2012). These characteristics support an exploration of the GAS and response processes, as therapists and clients are engaged in goal-directed behaviours which include processes that are primarily nonlinear (setting goals), but also include actions that are potentially linear (scaling goals). Therefore, applying APM provides an exploratory vehicle to investigate what response processes emerge when therapist-users engage with the GAS during dyadic interactions.

Gathering response process data of measures where content is formed by users and are used in dyadic contexts has, to my knowledge, never been explored. As such, the purpose of this study is to understand whether the APM protocol supports access to how therapists construct and

act on the use of the GAS in a therapeutic goal-setting interview with a client actor. By using an explanatory approach to validity (Zumbo, 2009, 2017a), this study seeks to address the following questions related to response processes and engagement with the GAS: (a) What response processes can be identified through the APM protocol when therapists are setting and scaling goals using the GAS with a client-actor? (b) How do identified response processes correspond to the procedures of goal setting and goal scaling during engagement with the GAS and a client-actor?

3.2 Methods

3.2.1 Protocol

As response processes have not been investigated in dyadic contexts, APM provided a method to facilitate an in-depth examination of the interaction between a client-actor and therapist-participant while using the GAS. APM was used to gather response process data from the perspective of a therapist, in the context of a conversation with a client. Figure 3.1 outlines the process of data collection and analyses using the APM, which begins with an interview and continues through initial analyses and a final interview that seeks feedback from the therapist.

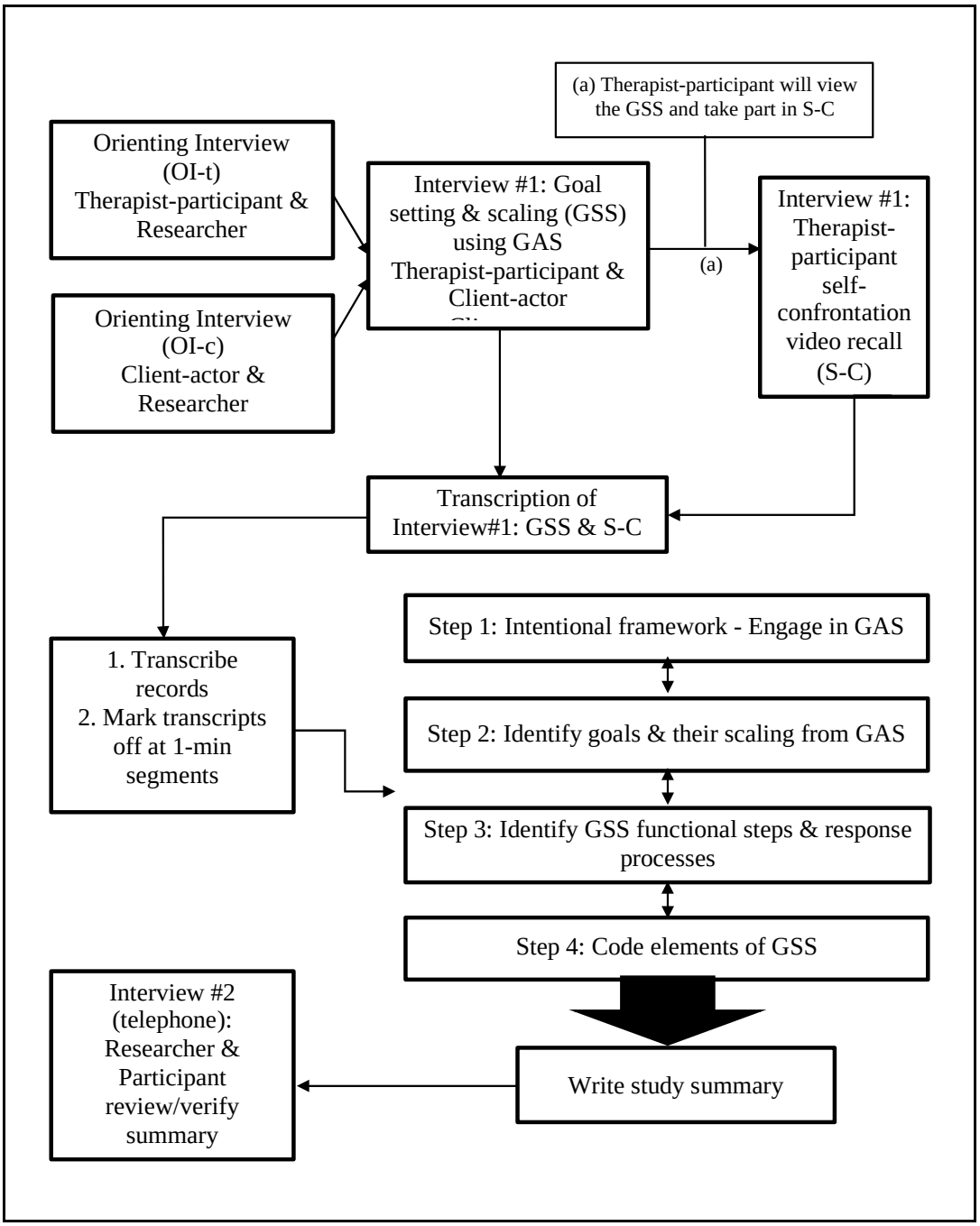

Figure 3.1 Data collection protocol & analysis procedures with therapist-participants

3.2.2 Participants

As Occupational Therapists are accustomed to assessing and evaluating goals in their professional practice, a sample of 7 therapist-participants were part of this qualitative investigation. Two actors, both male, were hired to role-play a client with a mental illness who is

living in the community and meeting with an Occupational Therapist for the first time to determine relevant goals. One client-actor interacted with the majority of therapists (n=6), while the other client-actor interacted with only one therapist.

Therapists were recruited by advertisements through the College of Occupational Therapists of British Columbia and social media, such as Facebook. To determine eligibility for recruitment, all therapists answered 'yes' to the following questions during a telephone screening: (1) Have you worked with individuals with mental illness, in any capacity? (2) Have you set goals with clients? (3) Have you used any goal-setting measure before? Therapists were all registered to practice Occupational Therapy with the College of Occupational Therapists British Columbia. Each therapist was given a $10 (CDN) parking voucher to compensate for their participation.

Therapist ranged in age from 26 to 61 years with a median age of 36.0 years. The median years of experience among therapists was 4.0 years and ranged from 3 months to 35 years. Five therapists were women and two were men. Therapists worked in both public and private practice, with children, youth or adults in the Lower Mainland of British Columbia. All therapists had current or prior experience working with adults with mental illness and using a goal-setting measure when working with clients.

The client-actors hired for the study met the following qualifications: (a) at least 1 year of experience acting on-campus or off-campus, (b) prior success with psychologically complicated roles, (c) experience and ability with improvisation, and (d) enthusiasm and interest. The client-actors were provided with a case scenario that outlined specific characteristics of the role and role-played a client with a mental illness who is interacting with an Occupational Therapist in an interview (see Appendix D for case scenario). A defined character was expected so there was

consistency in responses (i.e. interests of character, values, symptoms – current and past, and other details) between participants. The client-actor was paid $25 per session.

3.2.3 Procedure

Approval was obtained by the behavioural research ethics board at UBC (Ethics approval: H18-01915) for the protocol in this study. If therapists met eligibility requirements they were invited to partake in a video-recorded session at the university. At this time, therapists were also provided with the following to read over via email: (a) GAS instructions (cf. Kiresuk, Smith, & Cardillo, 1994) (Appendix A), (b) sample GAS form on which goals are written (Appendix B) and (c) a consent form. Upon arrival for participation, therapists signed the consent form and were provided with a hard copy of the GAS instructions, GAS form to write goals, a case scenario outlining information about the client-actor (Appendix D) and a blank notepad.

3.2.3.1 Therapist and client interaction

The orienting interview involved a brief introduction between the researcher, the therapist-participant and client-actor. During this time, each therapist was instructed to use the GAS to set and scale at least 1 goal with the client during a 30 minute video-recorded conversation. Therapists were told they were meeting with the client for the first time and can ask questions to probe for more details as they would clinically to get the information they needed; and were aware they were interacting with a client-actor. During participation, therapists were escorted to a room furnished with two chairs that were angled towards each other with a small table in front of the chairs. A computer with a camera was placed on another table in front of this set-up so the conversation between each therapist and client could be video recorded.

There was no structure for this conversation and therapists were able to ask any questions they wished to the client to use the GAS and determine at least one goal.

At the start of the APM procedure, the researcher briefly oriented the therapist and client to the settings and introduced them to each other. The researcher then left the room while the therapist directed the conversation and the use of the GAS, which was video recorded. After the conversation, the researcher conducted a brief debriefing interview with the client-actor, and two questions were asked. The questions were: (1) Did anything in the interview positively or negatively influence your ability to role-play the character? and (2) Is there anything else you would like to share about your experience? Immediately after this brief debrief with the client-actor, a self-confrontation interview with the therapist-participant occurred.

3.2.3.2 Self-confrontation procedure

To access the internal processes and social meaning of therapist's engagement with the GAS, a self-confrontation interview was conducted. After the 30 min conversation between the therapist-participant and client-actor, each therapist engaged in a self-confrontation interview with the researcher that was also video-recorded. The self-confrontation procedure included equipment to facilitate video play back of the goal setting and goal scaling conversation they just had with the client using the GAS. In this portion of the procedure, the researcher stopped the video-recording at approximately 1 to 2-min intervals and asked therapists to describe their thoughts and feelings during each segment. The questions "what were you thinking?" or "what were you feeling?" were asked during each segment to gather information about therapists' internal processes. Immediately after the self-confrontation a final debrief occurred. As described in the analysis below, the data from the conversation and self-confrontation resulted in a narrative of each therapist's use of the GAS.

3.2.4 Data Analysis

Both the conversation and self-confrontation interviews were transcribed verbatim. The data collected during the conversation and self-confrontation interview were analyzed to determine the overall goal for each therapist and to identify similarities between each case. This information provided insight into how therapists engaged with the GAS and the different ways response processes, such as cognitive, emotional, motivational or behavioural aspects that emerged.

3.2.4.1 Initial analysis

After the conversation was transcribed and marked at the same 1-2 minute segments as were used in the self-confrontation interview, the data were analyzed. The analysis process involved both a top-down and bottom-up approach to move from description to organization (Young et al., 2005). In the top-down procedure, the transcripts were read and video dialogues were watched to understand what the therapist and client were doing. This process moved from a general understanding of goal-setting using the GAS, to functional steps and how the GAS was used, and then understanding specific response processes. This process produced an overall goal or intentional framework for the therapist-client joint action. In the bottom-up procedure, the analyses move from understanding various elements of speech in the conversation to functional steps and then to goals (Wall et al., 2016). In this process, each turn of speech or *element* is coded or labeled using a pre-existing list of behaviours (e.g., "asks a question"; "describes self"). The elements are then linked to the data from the self-confrontation interview to integrate the and therapists' internal processes with the functional steps, or understand the ways therapists aimed to reach their overall goal. All steps are iterative to understand the *how* and *why* therapists engaged with the GAS to set and scale goals. Thus, all steps alternated between reviewing

transcripts in the conversation and self-confrontation, considering the intentional framework, examining the goals that are set and scaled as well as the functional steps required to set and scale goals, and then identifying the response processes that have emerged. Additionally, the GAS form was examined to see how therapists used the GAS and to understand how the GAS instructions were interpreted. The GAS form also provided some insight into therapist's motivation to engage with the tool. To verify the interpretation of the analysis, a narrative summary was written in lay language, and checked with each therapist-participant via telephone to ensure the summary accurately reflected their experience. This narrative described the therapist's goals and functional steps and included direct quotes from each therapist.

3.2.4.2 Within and cross case analysis.

Following the initial analysis, the data from the interview and evidence of negotiations or joint efforts were evaluated by members of the research team. For the within case analyses, all data were examined to identify categories and themes in response processes and patterns for each therapist. A case document was produced to summarize each therapist's intentional framework and goals of their conversation with the client, how the GAS was used, strategies, behaviours and response processes observed, and overall impressions. Next, information from how the GAS was used along with corresponding quotes were compiled, and similar statements were combined to form categories and then condensed into themes based on different response processes.

In the cross-case analyses, an instrumental case study approach using APM was used to guide an in-depth examination (Stake, 2005). The purpose of this approach is to describe the phenomenon of how therapists acted on and constructed the GAS across cases. The narratives produced were read and coded based on similarities and differences. Similar codes were grouped together to form clusters.

3.2.4.3 Trustworthiness

Trustworthiness of the findings was ensured through a number of procedures. An audit trail was created which included memos that were written after each interview, as well as detailed records of data collection and analyses. Each therapist-participant had opportunities to confirm processes and interpretations, such as: (a) in the self-confrontation procedure, which occurred immediately following the interview with the client, and (b) by phone, once the narrative summary was provided to therapists. The narratives from the interviews were checked with each therapist and no changes were requested. Narratives were also reviewed by all researchers. A debriefing interview was also conducted with each therapist and client-actor immediately following the self-confrontation procedure, which asked about their experience and whether they wanted to share any other information. During the analyses, data was coded and verified by more than one researcher and all researchers participated in the discussion of within and cross-case analyses.

3.3 Findings

This study investigated what response processes can be identified and how response processes correspond to the use of the GAS, as therapists act on and construct the GAS in a therapeutic interview. Engagement with the GAS was a process and the cross-case data analysis revealed that therapists' goal-directed actions involved two main foci. All seven therapist-participants shifted between focusing on *negotiating goals for goal-setting* and *formulating goals for the GAS*, where the shifts in foci involved bringing one or the other to the foreground (see Figure 3.2). When therapists were focused on *negotiating goals for goal-setting*, they were learning more about the client and also discussing possible goals with the client, and when focused on *formulating goals for the GAS*, therapists developed specific goals for entry into the

GAS. At the start of the interview, therapists started off negotiating goals for goal-setting and then proceeded towards formulating goals for the GAS. All therapists except one shifted their focus throughout the interview and as the interview progressed, the shifts between foci became less distinct. The foci were informed by various resources that therapists drew from, as well as their understanding of the term *goal*. All interactions between the client and the GAS provided response processes data. Cognitive and emotional response processes were most apparent as therapists articulated their thoughts and feelings during the self-confrontation procedure; however, engagement with the GAS and the client were also reflective of behavioural and motivational response processes. The findings described here show how therapists' act on and construct the GAS, as well as the different ways in which response processes emerged during goal-setting and goal-scaling. Therapist names below are pseudonyms.

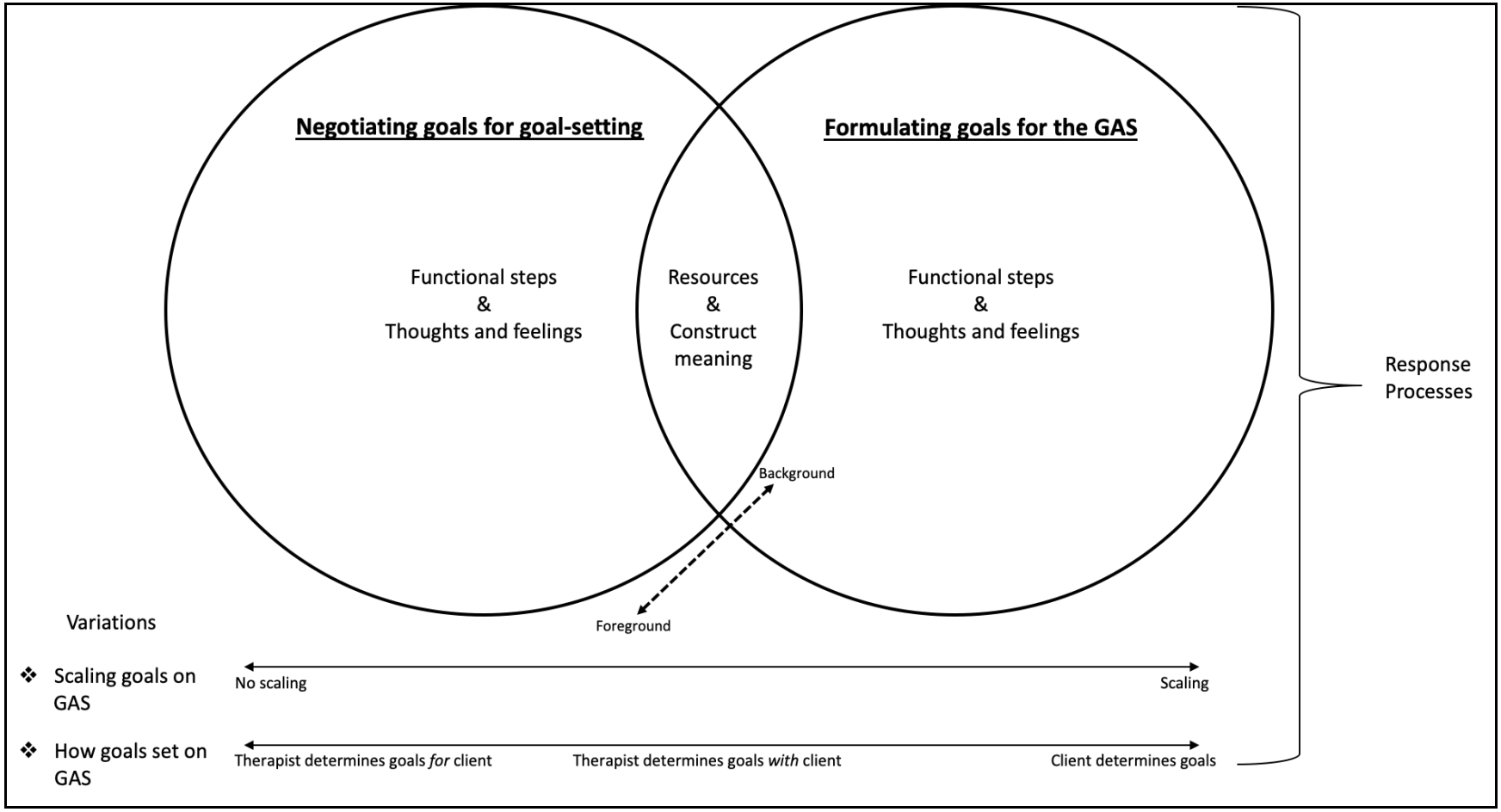

Figure 2.2 Engagement with the Goal Attainment Scaling measure

3.3.1 Negotiating goals for goal-setting

All therapists first focused on negotiating goals in the therapeutic interview to begin the process of engagement with the GAS. Consistent with education in occupational therapy and client-centered practice (Sumsion, 2000) all seven therapists began the process of negotiating goals for goal-setting by finding out more about the client, developing a rapport and attempting to understand the client and their reasons for exploring goals at this time - such as their illness history and barriers to possible goals. However, the findings presented here are directed towards therapists' engagement with the GAS. When therapists were focused on negotiating goals for goal-setting, they engaged in a conversation with the client and discussed what goals the client would be interested in pursuing.

As therapists brought negotiating goals for goal-setting to the foreground, all seven therapists participated in the functional step of exploring possible goals the client would like to work on. One therapist, Anna, was aiming to understand how the client was managing tasks at home near the start of their conversation. When asked in the self-confrontation, what she was thinking in this part of the conversation, she stated, "...in that moment like, initial goal-setting, I'm just kinda exploring like, yeah, what are possible goals." Similarly, another therapist, Olivia, was discussing things the client likes to do or used to do. When asked in the self-confrontation about her thoughts during this functional step, she explained, "...I was definitely trying to elicit some concrete ideas for goals to work on with him."

When focused on negotiating goals for goal-setting, all therapists employed their previous skills and training in their interactions with the client. Seema explained in the self-confrontation that she was, "...reverting back to the way I would set goals with clients and break it down." She elaborated in this self-confrontation that, "I was doing my schtick about goal-

setting. I was feeling I was the therapist and this is what I've done several times over my career." For all but one therapist, engaging in this initial process was also part of their engagement with the GAS. One therapist, Matt, remained solely focused on negotiating goals for goal-setting, and did not proceed towards formulating goals for the GAS. His sole focus became evident at the end of the conversation as he was summarizing the goals that were discussed and starting to wrap up the session. In this segment of the interview, Matt described some possibilities for the client by asking for clarification, stating a plan and also encouraging the client.

> Matt: Um, there are some things here that I would like to review. Um, so, I will write it down for you, is that ok?
>
> Client: Sure...
>
> Matt: ...we will just review some of the things that we discussed so completing a draft resume would be one of the things....
>
> Client: Right...
>
> Matt: And I will send you some samples, there might be some questions that you have. Feel free to ask and you might just select a sample and use that template to put your information in and then we can review it next session. Ok?
>
> Client: Yeah, ok.

As the interview segment continues, Matt looked down a couple times at the GAS form he was holding but does not write on it. During the self-confrontation interview, when Matt viewed this segment of the conversation described above, he explained, "I knew that I was setting goals, but I knew that I wasn't following the criteria. And I made a decision that I wasn't going to interrupt flow..." Altogether, his conversation focused on determining goals with the client, as identified by this therapist's overall goal for the conversation (see Table 3.1). The action and motivation to

not use the GAS was a response process that was influenced by the cognitive demands on the assessment. It was not uncommon for therapists to explain doing multiple things when engaging with this measure and as Matt aptly described in the self-confrontation, he was feeling, "...like we had a lot to work on and I didn't want to add anything else." However, the remaining six therapists who engaged with the GAS described negotiating goals for goal-setting as including the need to think ahead to prepare for using the GAS, and this was also evident in the intentional frameworks for their conversation (Table 3.1). Cindy expressed this notion while she was negotiating a preliminary goal with the client and described in the self-confrontation, "I was thinking about how to put that onto the...into the Goal Attainment Scale." The continuous process of engaging with the client and planning to use the GAS measure came before formulating goals for the GAS.

Table 3.1 Intentional framework while using GAS

Participant	Overall goal or Intentional Framework while using GAS
Seema	Setting up goals that are client centered & using the GAS
Matt	Determining goals with client-actor
Carlos	Using the GAS to determine goals with client-actor
Anna	Using the GAS to determine goals with client-actor
Cindy	Determine how to set goals using the GAS and with client-actor
Olivia	Use the GAS so client-actor can set and scale goals
Maya	Use the GAS in a controlled research setting to set and scales goals for client-actor

3.3.2 Formulating goals for the GAS

As therapists shifted their focus from negotiating goals, they brought forward their focus to formulate goals for the GAS. There were several ways therapists operationalized and spent time with each foci. In some instances, therapists negotiated goals for goal-setting and then

shifted their focus entirely to formulate goals for the GAS. For example, Seema spent the majority of the conversation on negotiating goals, and then spent the last bit of the conversation on formulating goals for the GAS. However, for other therapists their focus shifted throughout the conversation, particularly when more than one goal was set. When therapists did shift focus and brought forward formulating goals for the GAS, their actions could be understood through manifest behaviours and internal processes. These analyses of action include various levels of action such as functional steps and elements, all of which contributed to the meaning behind their actions.

3.3.2.1 Planning and prioritizing goals for the GAS

Moving from negotiating goals, therapists started to plan and prioritize goals either with the client or for the client, in anticipation of recording goals on the GAS measure. Four participants mentioned the process of planning and three identified prioritizing these goals. Carlos identified both, and in one functional step he acknowledged the client's goals and also prioritized them. During the self-confrontation interview, he explained that, "I wanted to make sure he...knew kinda what I meant by priorities and what is most important right now." Additionally, as Carlos was determining the details of one goal with the client, he explained thinking ahead towards use of the GAS in the self-confrontation, "I was trying to kinda plan ahead..." Olivia was trying to facilitate the client to articulate some goals and her thoughts during the self-confrontation illustrate the complexity of the goal-setting process for the GAS. She stated, "...especially with the GAS, there is quite a lot of information that you are including here...A snapshot of them is really easy to put down but then you have got a lot of information, a lot of planning that has to go on paper." Prioritizing and planning involved therapists thinking forward about how to employ the measure and also how to scale goals on the GAS.

3.3.2.2 Determining how to use the GAS measure

As five therapists were conversing with the client about goal-setting and goal-scaling, they were also thinking about how to use the GAS measure and follow the instructions identified in Kiresuk et al. (1994, p.7) (see Appendix A). During one conversation, Cindy was determining a preliminary goal with the client and introduced the GAS to the client. She explained, "We are going to use the Goal Attainment Scale today and so this just kinda helps us see where we are at with attaining your goal. So, if we think about the goal...related to employment or school." In the self-confrontation, Cindy explained her thought process for this moment noting, "...we have a goal here. So, I was thinking about how to put that onto the Goal Attainment Scale." Similarly, in this interview excerpt, Seema asked the client about possible goals and at the same time wondered how to put the information down on the GAS form:

> Seema: ...what would be reasonable – just wondering, can you come with some small goal, for yourself, around cooking?... what would that look like?
>
> Client: What would a goal be around that?
>
> Seema: Yeah. And it can go from anything like zero to trying some new lessons, to *<gestures with right arm out to right>*...what would it look like? If you worked in the food industry?
>
> Client: I guess one thing that occurs to me is that I have a friend whose cousin owns a restaurant.

When asked about her internal processes for this segment, Seema stated, "that's what I was kinda thinking...how do I use this form." As therapists were determining how to use the measure, they were also thinking ahead towards engagement with the GAS.

3.3.2.2.1 Thinking ahead and scaling goals

As part of determining how to use the GAS measure, four therapists identified prospective thinking while they focused on formulating goals for the GAS. In one conversation, Olivia set a goal with the client to go to the employment resource center. In this segment of the interview Olivia looked at the GAS form on the table in front of her and prepared to write; she stated in the interview to the client, "So, as a first step towards that. As an expected level of outcome, here I will just put down meeting with the counsellor...uh, meeting with the employment resource." During the self-confrontation she explained her thought process during this segment and stated, "...I was just thinking how are we going to do this, and as we are writing the expected level of outcome, how are we going to do the minus and the plus." The internal process highlighted in this example was shared by others - more specifically, among the six therapists that scaled goals, they all wondered how best goals could be scaled.

Scaling goals was an aspect of the GAS that therapists negotiated in the midst of their conversation and goal-setting with the client. For example, as Carlos was talking with the client, he asked, "Maybe we can talk then about...in terms of gaining employment...maybe we can talk more about full-time, part-time....probably thinking that best case scenario is a full-time job?" In the self-confrontation, he explained that during this segment, "I was thinking, my biggest concern was scaling these goals. So, I was basically thinking of ways to scale it." Similarly, during the self-confrontation, Anna explained some of her thoughts as she was setting a goal that was much more than expected with the client. She stated, "I was thinking that maybe with the way it is scaled, it doesn't have to be signing up for 1 class than 2 classes then 3. It doesn't have to be ratio like that but I guess it could be something that is really much more than expected."

3.3.2.3 Decision making strategies

Certain functional steps, as well as thoughts and feelings reflected particular decision-making strategies therapists used while engaging with the GAS. In particular, therapists used different techniques to shift decision making towards the client. As Carlos was guiding the client through the GAS some elements of their conversation were to prioritize goals, acknowledge the client, ask for the client's opinions, gather information, and also state a plan. When watching this segment of the interview during the self-confrontation, Carlos mentioned:

> I was feeling like, ok, how do I guide this process but how do I make sure he is, that these are his goals. Because it was me writing it down, it felt very much like I feel like, um, I almost wanted to give him the piece of paper and be like, 'ok, you write them down'...

Six participants suggested goals to the client or discussed goals with the client as a joint endeavour. In the following interview segment, Anna started to talk to the client about some goals he may have:

> Anna: I think I heard, possibly getting back into some sort of work, and maybe some fun activities. It sounds like you really enjoy cooking, and uh, exercising before.
>
> Client: Uh-huh.
>
> Anna: Any...I guess...is there anything else you would like to work towards, apart from the things that we talked about?
>
> Client: I don't think so, I think that kinda covers it.
>
> Anna: Ok. Is there one that takes priority over another?
>
> Client: Hmmm. You know, I guess work. I guess finding...figuring out what I want to do with work.

During the self-confrontation, Anna explained how she was trying to shift decision making towards the client, "...just in that moment I was just trying to point that out to help him realize that he's actually the one initiating all of this and not just some therapist telling him what he has to do." Therapists discussed how to use the GAS while practicing in a client-centered manner and three therapists articulated trying to get the client to set and scale goals to ensure they captured the client's desires.

3.3.2.3.1 Intuitive cues while engaging with the GAS

As part of the process of formulating goals for the GAS, therapists also drew from intuitive cues to help make decisions during engagement with the GAS. Specifically, two therapists identified relying on their intuition as an aspect that influenced their decisions and what goals they would set with the client. In this example, Maya suggested a goal to the client based on a hunch:

> Maya: One of the purposes of us meeting today, is to set a goal or two. We will see how it goes. And I think it would be interesting to chat about the fact of you going back to school. So, now that I've kinda told you a bit about these fancy aptitude tests, um do you think that is something you would be interested in taking?
>
> Client: I think so...

When asked how that goal came about in the verbal recall, Maya explains, "I just had a gut feeling that he didn't have any specific area that he was hoping to go back into. And knowing that that was the first thing to kinda attack." Cindy used a similar strategy to determine the client's level of interest as she was suggesting a goal related to the client's family and activities for his kids. As she was discussing her thoughts and feelings, and where she was getting some of her information to pursue this particular goal, she explained, "...nonverbal cues and even the tone

of his voice and just kinda how um, he..how he responded you know like, that might be he is not as keen as when we were talking earlier about school or employment."

3.3.2.4 Therapist impressions about engagement with the GAS

Therapists voiced various reactions towards their engagement with the GAS. Their thoughts and feelings about the measure varied and reflected a range of opinions. Three therapists voiced their frustration or dislike of the GAS measure. For Maya, scaling was stressful and she explained that during the self-confrontation, "...I was very stressed about trying to find timeframes that would fit in each category." Similarly, Seema stated during the self-confrontation, "I think it is my frustration with the scale" and that "I didn't like the form."

It was also apparent that therapists were uncomfortable and/or avoided scaling the lower levels of the GAS measure; four therapists expressed discomfort with the lower levels of the scale. Seema did not scale any goals at the less than expected levels. For Anna, as she was setting a goal that was at a less than expected outcome level with the client, she explained in the self-confrontation, "I was thinking oh, that's kinda hard conversation to have of setting, of asking what it would look like if you didn't quite achieve your expected outcome." Likewise, when Maya brought up the lower levels of expected outcome for one goal to the client, she described in the self-confrontation:

> ...knowing that I had to cover these areas, that made me feel uncomfortable...And I feel that when you talk about the people doing somewhat less than expected and much less than expected with this clientele in particular, it can maybe sometimes damage your ability to build rapport because these people who have depression are used to...uh, a number of people who have depression are used to not meeting goals.

On the other hand, two therapists voiced their appreciation for the GAS. Olivia was determining the expected level of outcome for a goal and how to scale this goal and she explained in the self-confrontation, "I could really see how this could be useful" and "...it gives me a sense of next week of where he is at towards this goal, how committed is he. If he is at minus two for everything, it gives us a starting point to talk about." During the debrief, when Anna was asked if she had any final reflections about using the GAS, she explained how this would be useful for her current work, "I was thinking about my current work actually and thinking that this is actually a pretty good tool because I like how, um, each rating is tied to a specific outcome..."

3.3.2.5 Following instructions for use of the GAS

Although therapists engaged with the GAS and were provided both a sample form and instructions (see Appendices B & C), no therapists used the GAS as it was intended. Whether intentional or unintentional, therapists voiced changes they made to the GAS and how they adapted the GAS to work for them. In this interview excerpt, Seema was confirming the first step of a goal set with the client and voiced uncertainty out loud about how to write down the goal they set:

Seema: Exploring cooking...ok. As a career option *<writing on GAS form>*

Client: Yeah.

Seema: And the first step of that is discussing the idea with your wife *<turns head to look down at GAS form but does not write>*

Client: Yes, and then the next step after that would be calling my friend

Seema: Ok...I'm not sure how to write this....and discuss with wife. <*glancing down at GAS form, pen positioned to write but does not write*> What are the chances of you talking to your wife about it over the next week – what do you think?

In this segment, Seema was breaking down the goal with the client and aimed to put some aspects of that onto the GAS form. She explained in the self-confrontation, "ok here's the goal and this is the step he was going to do and I was trying to put in the kinda stepped version..." Subsequently, Seema adapted how the GAS form was used and explained in the self-confrontation, "Um, but on the other hand we were getting the work done. Doesn't really matter what the form looks like..."

One therapist, Olivia, noticed a mistake with how she used the GAS assessment. She described her thought process in the self-confrontation when viewing a segment of the interview where she was discussing the less than and much less than expected outcomes with the client. In realizing her mistake she stated, "I made the mistake again of doing the plus 1 and plus two together before doing the 1s and the 2s." This was common and all therapists that scaled a goal fully completed the more than and much more than expected levels together (i.e. the +1 and +2), instead of first scaling the goal on 1s (i.e. +1 and -1) before the 2s (i.e. +2 and -2). Furthermore, upon examination of the GAS form, three therapists did not identify a goal title and instead put the identified goal as the expected level of outcome.

3.3.3. Resources influencing engagement with the GAS

A number of resources influenced therapists' use of the GAS and what they brought into the conversation with the client. During the debriefing interview, which followed both the interview and self-confrontation procedures, therapists were asked what they thought the term *goal* meant. They identified goals as something that was aimed towards the future (see Table

3.2). Six therapists explicitly identified a goal as something that one "works towards" or "wants to achieve," while one therapist discussed a goal more generally as being "intentional and purposeful."

In addition to the identified conceptions of a goal, therapists drew from various resources that influenced setting goals and subsequently their engagement with the GAS. Therapists considered these resources in both foci. Four therapists specifically identified the SMART (Specific, Measurable, Assignable, Realistic and Time-related) goal technique (Doran, 1981), as a framework they considered while setting and scaling goals during engagement with the GAS. Four therapists also mentioned their previous experience with the Canadian Occupational Performance Measure (COPM), which is a similar goal-setting measure to the GAS (Law et al., 1990). Other approaches that informed therapists' conversations and engagement with the GAS included Cognitive Behavioural Therapy, which Seema noted in the debrief interview that she drew from to "provide a little bit of education but also promote the activation side." One therapist mentioned being influenced by the Model of Human Occupation (cf. Kielhofner, 2008), and another mentioned a general self-management framework as resources they drew upon.

Table 3.2 Therapist conceptions of term goal

Participant	**What does the term/word *goal* mean to you?**
Seema	"Well it is intentional and purposeful. It has got purpose and attention"
Matt	"Well it means, in the context of this profession, something that you are working towards."
Carlos	"... an outcome of what we are working towards...It is basically like the end result we are trying to achieve."
Anna	"...it means to me something to work towards. Something you want to achieve or get to."
Cindy	"something to work towards"
Olivia	"It means something that you are not currently, you don't currently have or are not currently doing but you would like to achieve"

Maya	"A goal is something that someone wants to achieve, that's what it means to me."

3.4 Discussion

The purpose of this study was to understand how therapists act on and construct the GAS and investigate the response processes that emerge. Therapists had two main foci while they were engaged with the GAS, which were negotiating goals for goal setting and formulating goals for the GAS. Engagement with this measure was highly contextualized and therapists brought resources, and various thoughts and feelings to this process. The data revealed that engagement with the GAS included engagement with the client.

3.4.1 Response processes

As indicated in the findings, therapists engaged with the GAS in various ways and engagement involved a number of mental processes and operations. The layers of analyses involved with APM highlighted response process evidence throughout several aspects of the analyses, such as elements, functional steps, and therapists' thoughts and feelings. Response processes evidence for the GAS emerged as therapists engaged with the GAS measure (e.g. goal-setting and goal-scaling) and interacted with the client-actor, as well as from situational factors that contributed to this engagement. The complexity of response processes was revealed in this study through dyadic interactions, the overlap and exchanges between identified response process aspects, as well as contextual factors that link the interactions to response process evidence for the GAS.

3.4.1.1 Dyadic interactions using the GAS

This study draws attention to a unique measure that is commonly used between two people. To understand an action by the therapist using the GAS, one also needs to understand

that the action and its meaning is not contained solely within the therapist. In the process of engaging with the GAS, the therapist and client are interdependent, and the therapist and client rely on one another (Saavedra, Earley, & Van Dyne, 1993). They must communicate and cooperate interdependently to complete the task of engagement with the GAS. As an example, communication between a therapist and client is unpredictable and each utterance by a therapist will lead to a reaction by the client (Schuwirth, 2014). As such, the therapist is acting towards the client, as well as the GAS, so engagement is not limited to just the therapist and the GAS. The shared nature of goal-setting indicates that response processes with the therapist and the GAS must also include interactions with the client (Levack et al., 2016). In testing situations, although previous examinations have focused on individual response processes, it is recognized that even when separated in time and space, the activities of test takers are linked to raters (Fox, 2003). This interrelationship is an activity that is "laden with intent—directed at and defined by the perceived "other" " (Fox, 2003, p. 22). As indicated in this study (see Figure 3.3), the therapist must first interact with the client in order to engage with the GAS, and it is all of these interactions which are representative of response processes.

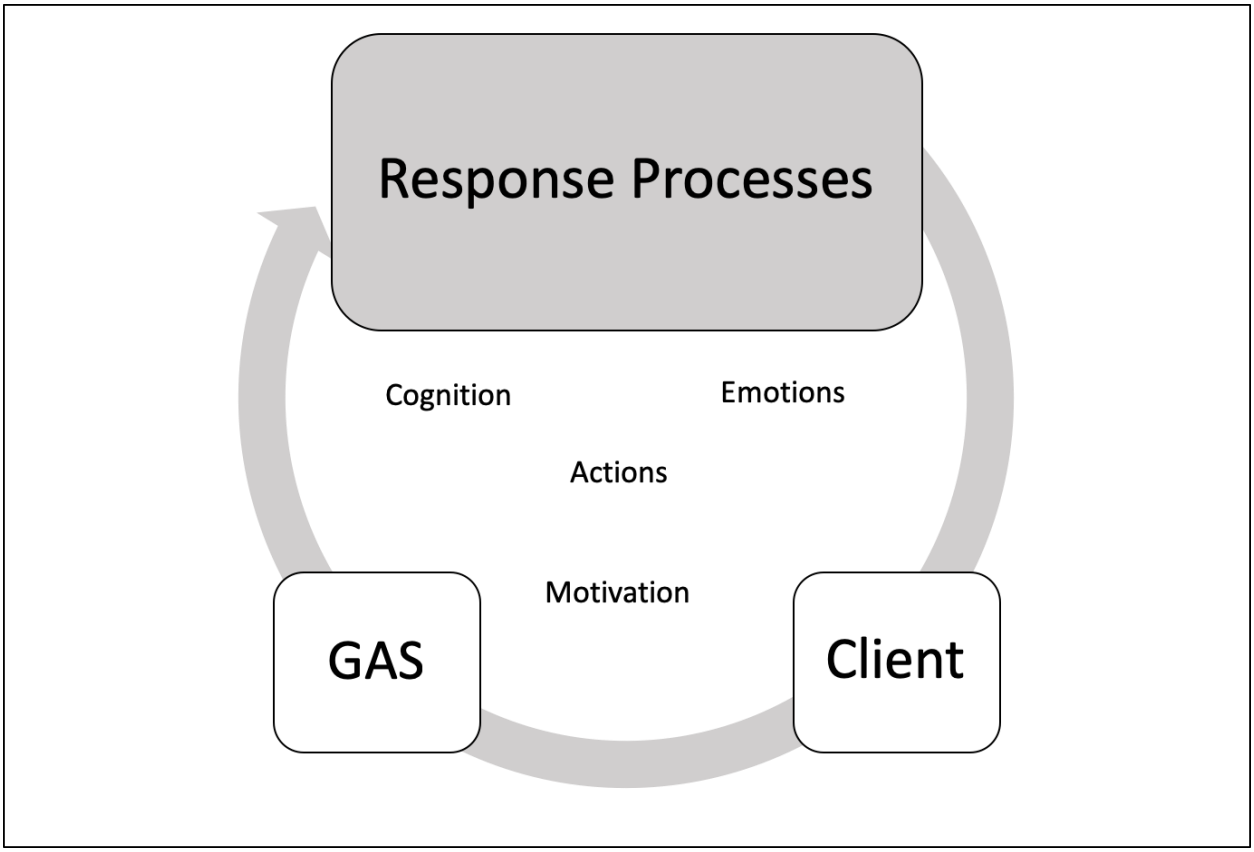

Figure 3.3 Response process model for the Goal Attainment Scaling measure

3.4.1.2 Overlap of response processes

This research moves beyond observed test responses, such as the final GAS score and goal content, to examine the processes that underlie these responses. Response process research tends to commonly report on cognitive processes, which are typically elucidated through cognitive interviewing, such as think-aloud interviews (Leighton, 2013). Using APM, this study shows that response processes are not restricted to one process and cannot be considered in isolation. For instance, therapist responses to the question, 'what were you thinking?' clearly indicate cognitive processes, but as shown in this study, therapist responses are not exclusive to this type of process. Researchers were able to observe actions that went alongside this cognitive response and also probe therapist's feelings when they reflected on that same moment. As

questions about therapists' thought processes and feelings coincided to similar moments watched during the self-confrontation, therapist articulations highlight that response processes overlap. Indeed, as users are engaged with a measure they are not only cognitive engaged but also motivationally and emotionally engaged (Launeanu & Hubley, 2017).

The overlap in response processes can be explained by first taking a look at individual processes. Cognitive processes governed how therapists engaged with the GAS measure and include interpretations of the GAS instructions, and utterances by the therapist about their interpretations of how the measure is used. Cognitive processes pertain to how participants understand the words (e.g. of a question) and their pragmatic meaning from the words (Schwarz, 1999). Reading and comprehending the GAS instructions and GAS form are examples of cognitive processes. However, as therapists' engagement with the GAS also involved interactions with a client, cognitive processes were influenced by emotional and motivational processes that helped steer their actions. As therapists were determining how to use the GAS, primarily a cognitive process, they may have also been feeling empathy towards the client or were more/less motivated to use the GAS as intended. As noted above, Anna explained that she was wanting to ensure the goal(s) set were what the client desired, and Seema adapted the GAS to work for her. Also, despite all therapists being instructed to *set and scale at least 1 goal*, Matt chose not to use the GAS. Instead, he chose to set three goals but did not scale any. These actions were influenced by his appraisal of importance and wanting to prioritize engagement with the client, which motivated his actions; however, they were also influenced by the cognitive demands of the measure. The overlap in response processes impacted how therapists used the GAS and whether it was used as intended.

Response processes related to therapists' actions showed how variably the GAS was used. Therapists' actions highlight that goals were set and scaled in a range of ways. Sometimes only the goal titles were noted or goal(s) were scaled partially or in full. Scaling of a goal also occurred quantitatively (i.e. timeframe and frequency) and qualitatively (i.e. level of achievement on one specific goal or different outcomes). The actions displayed indicate underlying processes that motivated these actions. It was evident that the GAS had a number of demand characteristics, as all therapists indicated multitasking, such as processing information from the present but also thinking ahead to next steps, all the while participating in a conversation with the client and also engaging with the GAS. Application of this measure in a conversation brought forth a number of response processes and highlighted how they are linked. This study shows that response processes do not just present through one type of process (e.g. cognition), since one process may be affected by others (e.g. cognitive processes may be impacted by emotional processes). The overlap and variability in response processes are also linked to numerous contextual factors.

3.4.1.3 Contextual factors

The contextual factors include the prospective nature of the GAS as well as the research environment and therapists' experience. During engagement with the GAS, therapists were constructing elements of this measure by considering the client's past experiences and projecting this information towards the future to set appropriate goals. A goal is a representation of a future possibility (Elliot & Fryer, 2008), and this definition represents how goals were determined on the GAS. Therapists needed to project forward with the client to consider goals that would be completed outside of the session at some future date and were also appropriate for the client. All goals discussed were with respect to the client's future and in anticipation that the client can and

will be able to complete the discussed goals, thus response processes were prospective. Even for self-reported items, responses from test takers are thought to be based on one's mental constructions about the future (Launeanu & Hubley, 2017). In addition to the future orientation of response processes, aspects such as the research environment (e.g. being video-taped, time constraints, and simulated client), resources utilized, familiarity with the GAS and managing the demands of the assessment were factors that played a role in how therapists engaged with the GAS. The variability demonstrated between how therapists engage with the GAS measure indicate differences at the individual level, the joint level between the individual and client, and also within the environment.

3.4.2 Action-project method and response processes

This study introduces a new method for studying response processes. APM provides of view of action that is in process and focuses on goal-directed processes over time that are jointly constructed (Wall et al., 2016). The procedures and detailed analyses involved with APM outline what thought processes, actions and feelings emerge as therapists act on and construct the GAS.

Through conversation, therapists were connecting information they collected from the client and integrating it with necessary elements of the GAS. Conversation is a joint action since it is the result of interactions between people (Young et al., 1999). Utilizing a method that enables a conversational view of response processes, acknowledges the interactive nature of engaging with a measure. Markus and Borsboom (2013) discuss testing as an interactive process between the test taker and test user. They use the metaphor "testing-as-conversation" (p. 256), which emphasizes a common understanding of the questions that are asked and answered, and they highlight the interpretative processes involved in test taking. There is clear alignment between actions resulting from test processes and actions captured from APM, since both

recognize that actions are interpretative. The various contextual layers uncovered provide insight into the construction of the GAS as dynamic, socially embedded and arising from the interplay between test takers and the test tasks; thus, individuals and their environments cannot be separated (Fox, 2003). APM allows for consideration of the context and exploration of how therapists act on and construct the GAS. The information gathered from this investigation can broaden the current definition of response processes to include one's goal-directed actions *and* intention.

Action theory, which forms the foundation of APM, sees human behaviour as not only goal-directed but also intentional, although not necessarily rational or linear (von Cranach et al., 1982). As evidenced by intentional frameworks (Table 3.1) and the sequence of actions determined from conversations between therapists and the client, most therapists' actions were geared towards setting goals and using the GAS. By delving into the processes and components behind those actions and intentions (e.g. manifest behaviours, internal processes, elements and functional steps), APM attempts to understand the social meaning to "yield findings that are consistent with people's interpretations of their experiences" (Young et al., 2005, p. 221). Validity also places central importance on interpretations and the use of scores (Hubley & Zumbo, 2011; Kane, 2013; O'Leary et al., 2017) - and herein lies the advantage of using APM as a vehicle to understand responses processes of a measure. APM enables a view of response processes that acknowledges engagement with the GAS is social in nature, where engagement operates at both individual and joint levels. Marrying APM, a method that includes goal-directed actions and intentions with response processes and an explanatory view of validity for testing, furthers our ability to investigate this source of validity evidence.

3.4.3 Construct validity evidence and response processes

Altogether, by investigating how therapists use the GAS, this study gathered key information about how and why therapists engage with measure. Previous validity research with the GAS has focused on its relation to other variables with no studies investigating validity evidence based on response processes (Shankar et al., *in press*). Investigating response processes links judgments about the content of a measure with consistencies in item responses, which is part of what Loevinger (1957) and Messick (1989b) describe as the substantive component of construct validity. The substantive component of validity investigates the, "context of measurement," (p. 661) which includes psychological theory and test behavior (Loevinger, 1957). These aspects contribute towards the overall construct validity of the GAS by providing information about how the goal construct is being measured. Within an explanation-based view of response processes, this study recognizes that response processes include contextual elements that are embedded within an ecological context (Zumbo, 2009). Applying this approach to testing situations highlights that involvement with any measure does not occur in isolation (Zumbo, 2017), but is influenced by the individual, client and environment. Furthermore, the testing situation can be seen as the interplay and exchange between: (a) characteristics of the user(s) and the social context, (b) the test taking processes that is occurring (e.g. completing the GAS), and (c) interactions over time (Fox, 2003). The aforementioned information situates therapists use of the GAS and helps to make sense of what they are doing in order to make valid inferences. As the meaning of the goal construct is linked to a range of tasks and situations to which it both generalizes and transfers to (Messick, 1994), response processes and substantive validity evidence provide information to help us understand how the goal construct is represented and measured in the GAS.

3.4.3.1 Goal construct and theory

This research points to the complexity of goals and demonstrates that engagement with the GAS confounds the following related goal constructs: goal intention, goal attainment and goal-setting. Each of these goal constructs have a different meaning and are attached to different goal behaviors. For instance, goal intentions do not automatically lead to goal attainment. As explained by Gollwitzer (1993) identifying goal intentions precede decisions to act on goals, and they suggest setting implementation intentions to commit an individual to specific plans related to their goal intention and executing this plan. Setting implementation intentions are more likely to result in goal achievement (Gollwitzer, 1993; Webb & Sheeran, 2007). In addition, goals change and several factors can contribute to the stability of goals, their level of endorsement and goal revisions (Fryer & Elliot, 2007). The GAS purports to measure the "degree of change" (Kiresuk et al., 1994, p.5) and as noted by findings in this study, perceptions of scale outcomes on the GAS may be related to a client's motivation or willingness to change. Consequently, it is necessary to understand whether the goals remain the same when the GAS is used at a later date to evaluate progress on the goals, and also examine whether the goals have shifted or changed.

In gathering validity evidence for this measure, it is of utmost importance to understand how this information contributes to our understanding of the construct measured by the GAS. Kiresuk et al. (1994) explained that, "the measurement procedure described here is a method of goal definition" (p. 445) which indicates that goals were originally defined through the method the GAS employs. In other words, *goal* as a psychological construct was not defined for the GAS. As noted by Elliot and Fryer (2008), how the goal construct is operationalized is inherently dependent on its definition, and ambiguity with the features of a goal will lead to operational variability - variability is demonstrated in this study. Although the goal construct for the GAS

was not defined during its initial development, this study points to the GAS as measuring goal intentions and employs the process of goal-setting. Evidence from this investigation indicate that therapists considered the word goal in similar ways by viewing goals as prospective in nature. In effect, the construct measured by the GAS during this initial interview are goal intentions. However, if one is setting goal intentions when goals are first determined on the GAS and then using the GAS at a later date to evaluate these goals, one is assuming that goal intentions lead to goal attainment. This assumption implies the goal construct changes (from goal intentions to goal attainment), which is not well supported in the literature (Webb & Sheeran, 2006). Even though these limitations with the goal construct will pose difficulty and confusion with operation of the measure, theory can improve and guide understanding of the construct (Koller, Levenson, & Glück, 2017).

Strong forms of construct validity evidence relate theoretical evidence with assumed relationships between the construct and a measure (Kane, 2013; Zumbo, 2009). Theory can help to specify boundaries and structure around the construct (Messick, 1989a, 1995). Although Kiresuk et al. (1994) do not make reference to a specific goal theory with regards to goal intentions and the GAS, there is mention of a goal-setting theory by Locke, Shaw, Saari, and Latham (1981). However, it is unclear how this goal-setting theory by Locke et al. (1981) is operationalized or applied in the GAS. In the current study, therapists made reference to a a technique for goal-writing called SMART, which stands for Specific, Measurable, Assignable, Realistic and Time-related (Doran, 1981). Although helpful in specifying goals, this technique does not link the *process* of goal-setting. In a Cochrane review of goal-setting strategies in an adult rehabilitation setting, Levack et al. (2016) found studies were limited in their use of theory for goal-setting. Similarly, Scobbie, Dixon, and Wyke (2011) in their review found the practice

of goal-setting is largely a-theoretical, yet theory is needed to guide and define mechanisms of action; they suggest some theories to consider for clinical applications. Altogether, the findings from this study indicate a need for theory (or explanatory model) to guide therapists' process-oriented thinking about the goal construct for the GAS.

3.4.3.2 Interpretations of the GAS

According to the Standards for Educational and Psychological Testing validity, "refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests," whereby the intended interpretations need evaluation (AERA et al., 2014, p. 11). An applied purpose of the GAS score is to evaluate change and the plausibility of this claim needs to be warranted through sufficient evidence (Kane, 2013). This study highlights that during initial use of the GAS measure, the GAS evaluates goal intentions. However, to evaluate change for an individual, the GAS measure rests on the assumption that one's initial goal intentions will lead to goal attainment. More generally, the GAS produces change scores and Kiresuk et al., (1994) indicate that the GAS has numerous applications (e.g. evaluate change within programs). From a measurement and validation standpoint, the obvious question that needs to be asked is 'what changes?' One also needs to consider how this change relates to our interpretations and the construct that one aims to measure, in order to demonstrate validity evidence for this tool.

The issues discussed here with regards to definition of the goal construct and the need for theory, raise the problem of potential or hidden invalidity within this measure (Flake & Fried, 2019). Of course, developing a measure comes with many challenges and concepts must be reduced to its component parts for their measurement. Nonetheless, the validity evidence based on response processes from this study indicate that at the very least, the GAS needs updating - to

reflect evidence about the nature of goal behaviours and for stronger measurement validation. The findings presented here provide a glimpse into some of the ways this measure can be improved for future use.

3.5 Limitations

Although this study offers a new approach to investigating response processes, there are limitations in the application of APM for this study. Due to the exploratory nature of this study, this investigation focused on therapist-participants' perceptions of their experience and did not explore the client-actors' perceptions. Future research can apply this methodology with real-life clients to understand more about the nature of joint processes involved in dyadic interactions, between both the therapist and client. A limitation pertaining to use of APM to examine response processes is that therapists' responses to questions such as 'what are you thinking?' and 'what are you feeling?' were not always exclusive to cognitive and emotional responses processes respectively. In particular, answers to the latter question about feelings were less distinct. Asking about feelings did not always provoke a response that pertained to one's feelings, or emotions and future studies can consider specifying this question when investigating response processes. Suggestions for modification include specifying that therapists reflect on feelings and emotions in the particular self-confrontation segment they are watching, such as 'what are you feeling in this segment?' and/or 'what emotions did you experience in this segment?'

3.6 Conclusions

This study contributes to the literature by examining how users engage with the GAS, which is a question that has never been investigated despite international use of this measure. By attempting to understand how therapists engage with the GAS, this study sheds light on validity

evidence based on responses processes during engagement with the GAS. How and why individuals respond the way they do on a measure is evidence that is central to measurement validation (Messick, 1989b, 1995; Zumbo, 2009). As demonstrated in this study, response processes are not limited to aspects that drive this process, but also include various interactions that surround what response processes emerge and how they emerge. I show that response processes cannot be isolated to interactions solely with the GAS as there was overlap between response processes when therapists engaged with the client and also with the GAS measure. Given advances in theory, measurement concepts, and conceptualizations of the goal construct, suggestions for future research to update the GAS are provided throughout. Altogether, this study broadens current views of response processes in a number of ways, which include: (a) examining a measure that involves two people engaging in a conversation, (b) introducing APM as an appropriate vehicle to investigate response processes and (c) expanding the definition of response processes to include one's intention. Response processes include several aspects that are nonlinear and connect various contextual factors, and through this investigation I provide insight into how these response processes dynamically unfold.

Chapter 4. Concluding thoughts - Connecting chapters, linking concepts and looking forward.

4.1 A recap of the problem and summary of findings

Chapter 1 outlined a unique problem in the extant psychometric and validity literature. The problem is that validation does not sufficiently consider tests in which a dyad is involved during completion of a measure. For my dissertation, I adapted the Action-Project Method (APM), a protocol that is commonly used to study joint processes, to gather validity evidence based on response processes. I apply APM to study the response processes that emerge during completion of the GAS, which is a measure that is well-known across disciplines and is commonly used dyadically. In Chapter 1, I explain that in the process of gathering validity evidence, information is being sought to build an argument for the intended interpretation and proposed use of the test. Thus, it is imperative to examine the assumptions and rationales behind the GAS score. I explain that substantive validity evidence is a key aspect in building a validity argument, as it examines the links between theory and response processes and its relation to the construct. As such, a key part of gathering response process evidence for the GAS is relating this validity evidence to the interpretation of scores and the goal construct for the GAS.

In Chapter 2, I start with a review of validation practices to understand how validity evidence has been gathered about the GAS measure. I use the unified validity framework to examine how validity evidence has been collected, understand what has been done and what is missing. I uncover several gaps in the available literature regarding how validity evidence has been gathered for this measure and its relation to the goal construct in the GAS. I discover that the term *goal* is used to refer to several aspects of the goal construct, such as goal attainment or

goal achievement, and goal-setting. The inability to identify a clear goal construct presents some difficulty with interpretations of the GAS and suggests the need for stronger connections to a goal theory before such inferences can be strengthened. I show that validity evidence pertaining to relations to other variables was over-represented in all articles, as most studies examining validity compared the GAS score with other measures. Overall, Chapter 2 illustrates a gap in validity evidence pertaining to response processes of the GAS, which is a key aspect to understand how the GAS measure is used between two individuals and also how validity evidence is collected for this measure. This review demonstrates that substantive validity evidence is largely absent, which opens an opportunity to investigate response processes for a measure that is used dyadically.

I follow-up from the information obtained in Chapter 2 to explore the verified gap in substantive validity evidence for the GAS. In Chapter 3, I adapt the APM to investigate response processes as two individuals engage with the GAS measure. By adapting APM to use for measurement purposes, I am able to explore how therapists act on and construct the content of the GAS and gather validity evidence based on response processes. Effectively, by applying APM in this context, I am able to investigate the joint nature of using this measure, as well as how the goal construct is represented. I demonstrate that APM is a feasible method to capture the interactive and goal-oriented nature of this measure, and show that APM has promise for collecting response process information with other measures. Response process information revealed how dyads interact with the GAS measure, the overlapping nature of response process, as well as contextual aspects that influence how the GAS is used. Using APM to investigate response processes also revealed the underlying assumptions about the goal construct in the GAS as attempting to measure goal intentions, goal attainment and goal-setting. From Chapter 3, I am

able to conclude that APM is a suitable and useful method to investigate response processes, and enables an opportunity for researchers to investigate beyond cognitive response processes, such as actions, emotions and motivation. Altogether, by introducing a novel method, this study demonstrates the complexity of response processes and how these processes dynamically unfold during completion of a measure.

4.2 Bringing the concepts and methods together

In this section, I aim to connect the concepts and methods that are used throughout this dissertation to clearly convey the advancements and innovation produced by this research. Starting with use of the GAS as a measure, this section explains why the GAS is a suitable measure from which to examine validity and why application of validity theory to the GAS is fitting. This initial section also explains possible sources of invalidity evidence for the GAS and how validity information for the GAS should be assessed. Next, I outline why validity theory and action theory, specifically APM, are well-matched. Finally, I conclude by explaining how applying APM as a new validation method for response processes can bring significant new insights for response process research and help steer the future trajectory of validity research.

4.2.1 Can validity be investigated for the GAS, and is the GAS valid?

The GAS is an atypical approach to gather score-based information and one may question its legitimacy as a measure. The GAS does not include specific content and uses a combination of approaches to gather information; it is comprised of both an interview and a rating scale to yield data to produce an eventual score. Interviews involve interactions between users, which obtain information about the respondent's specific problem areas, and a rating scale is used to judge expectations and performance. In moving from interviews to measurement, the consistency in behaviours or item responses needs to be justified (Loevinger, 1957; Messick, 1989b). As

noted by Messick (1989b, p.14), "measurement inferences are drawn from *scores*," which summarizes observed consistencies on a measure. Using the GAS means translating values from the rating scale to a final GAS score. Although the construction of the GAS measure is unlike more commonly used self-report formats, they both produce scores that are used for evaluation purposes. The term "test score" (p.5) is broadly used to refer to any observed consistency, and includes any means to document or observe consistent behaviours; it includes qualitative as well as quantitative summaries (Messick, 1989a). As such, what is validated is not the test itself, but the inferences about score meaning and interpretation, as well as the implications for actions that these interpretations entail (Messick, 1989a). In this context, the meaning of a measure is both context-specific and generalizable (Loevinger, 1957; Messick, 1989b). Thus, investigations of validity information for the GAS need to focus on the interpretations that are made from the GAS score and its score meaning, which according to developers of the GAS, represent the "degree of change" (Kiresuk et al., 1994, p.5). To recognize the significance of score meaning, Chapter 2 explains that the need for understanding the nature of the change in the GAS hinges on clarity of the construct. In Chapter 3, I reflect the idea of change back to the construct being measured and explain the implications of a changing or dual goal construct.

While regard for the GAS as a measure is reflected in its production and use of scores, an additional consideration is the application of validity theory towards this measure. Determining whether the GAS is an appropriate tool from which to apply current validity theory can be seen by previous studies investigating validity information. As shown in Chapter 2 and the numerous studies that have investigated validity evidence for the GAS, it is clear this is not a question of whether validity theory can be applied to the GAS, but rather how it is applied towards the GAS. While it is difficult to know the exact reasons why response processes have yet to be investigated

for the GAS, it is not uncommon that response process evidence is rarely presented in practice (Zumbo & Chan, 2014). Given the unusual format of the GAS, probing the ways in which individuals use the measure illuminates the underlying processes involved with the task, which is demonstrated in this dissertation.

Just as investigations of validity and application of validity theory are relevant to the GAS, it is also essential to consider how invalidity evidence, or threats may influence the results obtained from the GAS. Certainly, a lack of clarity about the goal construct in the GAS (e.g. no construct definition) opens up the possibility of other sources of invalidity, which can have negative social consequences (cf. Messick, 1989b). Two sources of invalidity that contribute to the consequences of legitimate test use and the soundness of the score meaning are construct underrepresentation and construct-irrelevant variance (Hubley & Zumbo, 2011; Messick, 1989b). With construct underrepresentation, measurement of the construct is too narrow and does not adequately encompass the full range of the construct a test is aiming to measure (Messick, 1995). Alternatively, construct-irrelevant variance pertains to measurement of a construct that is too broad (Messick, 1995). It contains variance that is associated with other constructs as well as method variance, which pertains to the test method used and its effect on responses that are irrelevant to the intended construct (Messick, 1995). According to Kane (2013, p.40), "limiting the test to one method of assessment for a broadly defined trait can lead to both underrepresentation of the trait and irrelevant method variance." He further describes ways to minimize this source of invalidity by ensuring strong construct validity, or inclusion of a theory that explains the plausibility of the assumed relationship between the test and the construct. Based on the investigations in this dissertation, the GAS appears to have potential sources of invalidity that relate to construct clarity, which needs further consideration to minimize the

negative consequences that can arise from use of this measure. While the positive consequences of using the GAS can lead to an accurate description of an individual's goal intentions, this dissertation points to several aspects of the measure that may lead to negative consequences during its legitimate use. An unintended or negative consequence of using the GAS may be that the goal(s) articulated during initial use of the GAS were later evaluated as unattainable, yet the individual's goals changed and hence different goals were achieved. Subsequent ramifications for whom the goals on the GAS reflect can be negative (e.g. perception that the individual lacks motivation), particularly if one uses the GAS solely as a measure of an individual's ability to attain goals, without consideration of relevant research on this topic (see Chapter 3). Therefore, issues related to potential sources of invalidity in the GAS, such as: the construct definition, a dual or changing goal construct, inconsistencies in how therapists used the measure, as well as variability in responding, can all help to illuminate areas for improvement in the GAS.

Potential sources of invalidity for the GAS can distort the meaning of the results and can place inferential limits on a measure (Zumbo, 2007b). That is, score inferences may be valid for some groups of users in some contexts and not for others (Zumbo, 2007b). Sources of invalidity can stem from the goal construct and related variance, as well as consequences arising from how individuals interact with the measure. As noted by Austin and Vancouver (1996) there are limitations when assessing goals through an individual's self-report, since they depend on what is accessible to an individual at a conscious level, as well as their awareness of goals. Thus, lack of alignment between the goals being measured and an individual's actual goals (and subsequent actions) can introduce invalidity in the measure of one's goals and influence the consequences of test use (Austin & Vancouver, 1996). Any assessment of goals using an individual's self-report will be under such a construct threat, especially since goals have the potential to change.

However, as noted in Chapter 3, ensuring a clear construct definition and score meaning can mitigate misinterpretations and help to minimize factors that can contribute to invalidity.

Taking into considertation the investigations and discussion in this dissertation about the validity evidence for the GAS, a legitimate question to ask is whether the GAS is a valid tool? The decision to designate a tool as either 'valid/not valid' implies validity is a fixed property, and it is a dichotomous label that is no longer recommended (Bandalos, 2017; Cizek, 2016; Zumbo, 2007a). As noted by the *Standards*, validation is a process and validity is based on the accumulated evidence to support the interpretations of a proposed use of a test (AERA et al., 2014). Consequently, what is validated is not the tool itself, but the interpretations that are made from the tool; hence, validity needs to be described along a continuum (Zumbo, 2007a). In the process of evaluating validity evidence for the GAS and proposing a new validation method to investigate response processes, this dissertation uncovered some critical inferential limits for the GAS. In particular, this dissertation highlighted that the validity evidence to support intended interpretations of this tool to measure the broad construct of goals, and/or the "degree of change" (Kiresuk et al., 1994, p.5) is not substantial and is incomplete. In terms of the proposed use of the GAS to assess the broad notion of goals and then understand the score as reflecting one's goal attainment, it is only realistic to say that there is insufficient evidence to support use of the tool in these ways. Given the validity evidence that exists mainly reflects the relation of the GAS to other variables (i.e. Chapter 2), this dissertation is the first to provide validity evidence related to response processes (i.e. Chapter 3) in order to look deeper into the actual use and interpretations of this measure.

4.2.2 APM, action theory and validity theory

Until now, the methods used to investigate response processes have been limited in their ability to provide a comprehensive view of the different ways response process emerge during engagement with a test. Applying APM in this dissertation enabled a discovery that exceeded initial expectations of the usefulness of this method. In particular, using APM in a measurement context revealed its: (a) utility towards joint measures, (b) ability to capture numerous response processes and (c) flexibility towards measures with nonstandard formats. Using APM highlighted the various contextual factors that need to be considered during interaction with a test, thereby enabling an expansive view into test-taking that is unmatched by conventional methods.

By investigating the substantive component of validity through the use of action theory, APM provides a way to more fully examine the "context of measurement" (Loevinger, 1957, p.661). As first noted by Loevinger (1957), in order to make appropriate inferences from a test, the test needs to be both representative of behaviours outside of the test and also relevant to the field of study. It is through this examination of the context that one gathers substantive validity evidence and can appropriately link both test behaviour and test theory to the construct being studied. This dissertation links the action of interacting with a test, a goal-directed action, with investigations of response processes for a goal measure. Specifically, APM, a method that is congruent with investigating goal-directed processes, is applied in a measurement context as a method to investigate the process of test taking with a measure of goals. Using APM for investigations of response processes affords a view of the context of measurement, including the setting and culture of testing. Furthermore, using APM to investigate response processes helps to derive explanations for responses on the GAS that provide an understanding of the reasons

individuals respond the way they do (e.g. Figure 3.2), instead of basic descriptions of their responses.

In all, APM provides a glance into the generative space, or the time between when the respondent is asked about a goal and when they provide a response that can be recorded on the GAS (cf. Zumbo, 2017a). In particular, APM helps to elucidate an understanding of the various response processes that are at play when one is interacting with the GAS (e.g. Figure 3.3). As a method conceptually grounded in contextual action theory (CAT), APM enables analyses, "that are open to constructing local and specific explanations without losing either a common language or a grounding in everyday experience" (Young, Marshall, & Valach, 2007, p.16). As APM draws from CAT, the three perspectives of action (manifest behaviours, internal processes and social meaning) provide data that align with investigations of response processes. The perspectives in this CAT-based methodology are not all shared by conventional methods to study response processes. Through video recordings of manifest behaviours, it is possible to see what an individual actually does when they interact with a measure, and this includes interactions between two individuals when using the GAS. By collecting information through self-confrontation interviews, the accompanying thoughts and feelings or cognitive and emotional processes are also elucidated. Investigating cognitive processes is the most common topic in response process research. However, the self-confrontation interview in APM includes emotions to more fully understand the processes that steer goal-directed actions and their meaning (Valach & Young, 2002). As shown in this dissertation, APM also enabled a glimpse into some non-cognitive factors (e.g. motivation, intuition) that may influence decision making during use of the GAS (cf. Young & Valach, 2008). Systematic observations combine both the video-recorded observations and self-confrontation interviews to provide an understanding of the subjective

experience or social meaning of a goal (Valach and Young, 2009; Young et al., 2018). Together these three perspectives complement the processes involved in interacting with a test. For instance, manifest behaviours provide information about how an individual associates with a measure, and the observable responses that are discussed or perhaps, documented on paper. As well, examining thoughts and feelings provides insight into how responses were generated. Finally, the social meaning brings together observations and interviews to gain a sense of what the experience of interacting with the GAS is about. As shown in this dissertation, APM revealed that the social meaning of interacting with the GAS involves much complexity, such as the variability in responses by test users, as well as a changing goal construct. The ability to document the social meaning and social construction of the GAS provided pertinent information about the construct the GAS purports to measure. Since the unified view of validity integrates all available validity information back to the construct, the combination of APM and the explanation-focused view of validation provided key explanations for understanding underlying construct assumptions for the GAS measure. As shown in this dissertation, APM can investigate response processes, but also greatly informs how this information relates back to the construct a tool intends to measure. The rich information provided by APM can serve to inform the future of validity research and how individuals interact with tests.

4.2.3 New frontiers for response process research using APM

By integrating an ecological perspective of item and test performance, APM as a new method to investigate response processes explicitly moves validity research from an *in vitro* view of testing to an *in vivo* view (Zumbo, 2015; Zumbo et al., 2017a). The shift towards an *in vivo* view means a better ability at capturing the context, and the "process of interaction and social embeddedness of the testing situation that inform and mediate individual response processes"

(Maddox & Zumbo, 2017, p.180). Specifically, APM enriches the contextual information that can be obtained during testing by collecting data from several action and process perspectives. As an example, APM enables insight into both affective and cognitive processes, instead of just one process at a time. Certainly, it embodies the saying 'more bang for your buck' in terms of the quantity and quality of response processes data that are collected. Since APM includes video-recordings and the unit of analyses is action, the method is well-suited for capturing any observable actions that are involved in testing situations. For instance, as Maddox (2017) revealed, capturing information on gesture can also provide insight into the features of interactions during testing. Similarly, other forms of response process data that rely on observations, such as reaction time are also plausible using APM.

Another innovative appeal for the application of APM in validity research, is its ability to get a deeper understanding of the construct a tool claims to evaluate. As shown in Chapter 3, APM facilitated an understanding for how the goal construct in the GAS was interpreted among users, and also how this construct may change when the GAS is used later to evaluate goals. In particular, APM helped to understand the dual goal construct (goal intention and goal attainment) embedded in use of the GAS; a feature that was not identified during its initial development (Kiresuk & Sherman, 1968; Kiresuk et al., 1994). Furthermore, as demonstrated in Chapter 3, examining how therapists interacted with the GAS provided an understanding about their priorities as they engaged in a conversation with the client-actor. APM facilitated an understanding of aspects such as building rapport or understanding possible barriers for a client, which were all elements that therapists considered while engaged with the GAS. These aspects provide an understanding of the demand characteristics of this joint measure, thereby providing

insight into how *well* the goal construct is being measured, as well as possible sources of invalidity.

4.3 Novel contributions

There are a number of novel contributions that arise from this dissertation and in this section, I identify four innovative features. The first contribution from this dissertation is examination of validity evidence for tools that do not follow a standard format. As the GAS has content that is formed by the respondent and/or users, it is unlike conventional measures which are typically in a self-report format and include pre-determined items. In this dissertation, I show that it is possible to study response processes for a unique measure such as the GAS. By using APM to gather substantive validity evidence, I demonstrate the importance of considering the context of measurement and how users engage with the GAS as critical parts of validation processes and building a strong validity argument. Thus, my dissertation provides a model from which to examine validity evidence for nonstandard types of tools.

A second novel contribution from this dissertation is a glance into multiple response processes during interaction with a test. Previous research in response processes has focused almost exclusively on cognitive approaches (Launeanu & Hubley, 2017), but as shown in Chapter 3 of this dissertation, APM allows for examination of response process that include emotions, actions and motivation. By exposing multiple response processes at once, Chapter 3 demonstrates how variably people construct and respond to the GAS, which also highlights potential sources of hidden invalidity for the measure. Furthermore, using APM revealed the overlapping nature of response processes, and how certain aspects steered therapists' responses to the measure. Applying APM for testing purposes provides an innovative look into response processes that is unmatched by other methods currently used to investigate this source of validity

evidence. APM highlights that the meaning of a test goes beyond observed responses since engagement with a test also corresponds with the realities that exist in the human world, such as the action of test-taking or having a conversation. Through APM, I show that the social meaning associated with testing is a dynamic relationship between users and the context of testing.

The third novel contribution that surfaces from this dissertation pertains to the construct measured by the GAS. Through an investigation of response processes of the GAS, this dissertation illustrates that the goal construct changes form during use of this measure. Specifically, in Chapter 3 I reveal the assumptions underlying the GAS, that the goal construct changes from its initial use, such that goals measured at the start evaluate goal intentions and later are assumed to measure their achievement. I also discuss the complications and limitations that arise from attempting to measure these dual constructs, particularly the assumption that the goals measured when the GAS is initially used are equivalent to the goals evaluated later using the GAS. A clearly defined construct is outlined as a basic and essential step during scale development that, when overlooked, can lead to complications with how a measure is used and the resulting inferences. Chapter 2 highlights that validity information gathered for this measure rarely includes a definition of the goal construct, which places constraints around how the tool is interpreted. This rudimentary step during scale development can minimize unintended consequences and help to ensure a measure is interpreted as intended.

The final and principal contribution that surfaces from this dissertation is the application of APM as a new method to study validity evidence based on response processes. Overall, applying a contextual, process-oriented method that explains actions is highly compatible with investigations of a source of validity evidence that is itself a process, and necessitate consideration and explanations of the test environment. There are two additional layers that add

congruence to the compatibility of APM for investigating response processes in this dissertation. Firstly, APM is a method that studies goal-directed behaviours and test-taking is a behaviour that is goal-directed, and secondly, the GAS is also a measure that aims to evaluate goals. By using a method (APM) that is congruent with what this dissertation studies (response processes), and what participants are constructing (GAS), there are multiple layers of congruence that substantiate the fit of APM for validity research. These layers of congruence add to the strength in determining APM as a suitable method to study response processes. Certainly, the innovative contribution of applying APM for validity research provides a new way forward for both validity research and the study of response processes.

4.3 Limitations

The chapters presented here offer some new perspectives about how response processes information can be collected. As in all research there are, of course, some limitations that must be considered. A foremost limitation arising from this study is in Chapter 3, where I introduce APM to investigate validity evidence based on response processes. As noted in that chapter, this study only investigated response processes from the perspective of the therapist. In doing so, I am able to infer the jointness of the GAS measure, as a shared process between the therapist and client but I am unable to provide direct evidence for it. Thus, a broader question that was purposefully controlled in this study is how joint processes, from the perspective of *both* users, influence test interpretations. Since this study was the first of its kind to apply APM in a measurement context, examining response processes from the perspective of one individual helped to determine if APM is a feasible method to investigate response processes.

Another limitation that arises from this dissertation is the reliance on the *Standards* as a framework to examine validity evidence (AERA et al., 2014). Although one may question why I

have relied on this resource, it is important to point out that the *Standards* (AERA et al., 2014) was jointly produced by three organizations (i.e. APA, AERA & NCME), who are leaders in the testing community. Despite the document being deemed as a "test standard" it is still not widely promulgated across disciplines. Although more and more researchers are becoming aware of the *Standards*, some of the studies reviewed in Chapter 2 may not have been aware of this resource. As a result, the change in language from considering validity as a property to its consideration as an integrated perspective of one's intended use has not yet been fully adopted. Nonetheless, demonstrating how the *Standards* can inform validation, as in this dissertation, can guide future researchers to use validity appropriate resources to support their research.

4.4 Implications for researchers and future considerations

Although some discussions about future applications are interspersed throughout Chapters 2 and 3, this section summarizes specific implications for researchers and ties them with areas for future consideration. This dissertation has a number of implications that can augment future research in the areas of validity, testing and scale development. As outlined in Chapter 2, there remains much confusion about the application of the unified view of validity in practice. By detailing how validation practices were investigated, I show that validity evidence needs to consider evidence outside of relations to other variables to provide evidence for what a tool claims to measure. The various sources of validity evidence are described and ways to improve validity evidence are discussed. Although the preferred validity approach is up to the researcher, I highlight that simply outlining what approach one chooses to use can aid interpretations of a measure, as well as the importance of relating that evidence back to what a tool claims to measure. This dissertation employs the *Standards* to guide my thinking of validity

and through its use, also demonstrates how it can be used to aid other researchers and their thinking about notions of validity.

By synthesizing information from a range of sources in Chapter 2, I illustrate the importance of looking in different areas, such as across disciplines and search engines, to find gaps in evidence. I show the importance of not relying on expected or famous sources of evidence for a measure, but delving into the literature to truly understand how a measure has been interpreted. The synthesis provided in Chapter 2 can also remind researchers about the necessity of a comprehensive review before decisions or judgments about the validity of a measure is made, as accumulated validity evidence can help determine the construct validity (or lack thereof) for a measure. As noted in Chapter 2, another implication for researchers is the need to state construct definitions to outline the conceptual meaning during tool development. Using theory to guide the definition of a construct at the start of tool development will carry forward implications towards its score meaning as well as its future use.

In Chapter 3, I use APM to investigate response processes for a dyadic measure and correspondingly demonstrate its usefulness in gathering response processes data. This study is the first to use APM for measurement purposes and demonstrates both the feasibility of this method for dyadic contexts and its value for response process research. In this investigation I highlight the importance of not overlooking a second person that may be involved in testing situations. From the variability demonstrated in Chapter 3 between individuals in the same profession, I show that interactions with the GAS are distinct, and researchers cannot assume performance will be the same across users. Although response process evidence for the GAS has been largely overlooked, I show in Chapter 3, the rich amount of information that APM reveals about how individuals interact with a measure and how joint processes can influence this

interaction. There are many opportunities for future research to apply this method and understand how measures are used between two individuals, as well as delving deeper into the response processes of the second person (e.g. the client). Certainly, researchers can use the methods described here and my findings as a model to guide how response process evidence can be retrieved using APM. My research also contributes to broadening perspectives of response processes in my third chapter by demonstrating how one's actions, emotions and motivation are also involved during engagement with a measure. Thus, researchers can utilize APM to investigate response processes with other measures because the APM goes beyond cognitive processes.

Altogether, the research that is presented in this dissertation highlights a number of points related to construct validity evidence for a measure and the importance of understanding the context of measurement. I demonstrate the need for substantive evidence and how building a sufficient validity argument contributes to evidence for what a measure claims to evaluate. This dissertation shows that response processes are rarely explored in the extant validity literature for the GAS. Certainly from a student- and patient-centered perspective, overlooking this evidence will only limit testing and ignore the diversity of test users if their perspectives are not considered. Applying APM for testing purposes provides an innovative look into response processes that is unmatched by other methods currently used to investigate this source of validity evidence. As a whole, this dissertation aimed to address a gap in validity evidence, but the research presented here also offers many opportunities for future research, in the realms of validity, testing, and scale development.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. [AERA, APA, & NCME] (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, *37*, 1–15.

Austin, J. T., & Vancouver, J. B. (1996). Goal constructs in psychology: Structure, process, and content. *Psychological Bulletin*, *120*, 338–375. https://doi.org/10.1037/0033-2909.120.3.338

Bandalos, D. L. (2017). Chapter 11: Validity. In *Measurement Theory and Applications for the Social Sciences*. (pp. 254–297). New York, NY: Guilford Press

Bandura, A. (1997). *Self efficacy: The exercise of control*. New York: W.H. Freeman.

Barry, A. E., Chaney, B., Piazza-Gardner, A. K., & Chavarria, E. A. (2014). Validity and Reliability Reporting Practices in the Field of Health Education and Behavior. *Health Education & Behavior*, *41*, 12–18. https://doi.org/10.1177/1090198113483139

Beaton, D. E., Bombardier, C., Katz, J. N., & Wright, J. G. (2001). A taxonomy for responsiveness. *Journal of Clinical Epidemiology*, *54*, 1204–1217.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Bouwens, S., van Heugten, C. M., & Verhey, F. R. J. (2009). The practical use of goal attainment scaling for people with acquired brain injury who receive cognitive rehabilitation. *Clinical Rehabilitation*, *23*, 310–320.

Bronfenbrenner, U. (1979). *The Ecology of Human Development*. United States of America: Harvard University Press.

Calsyn, R. J., & Davidson, W. S. (1978). Do we really want a program evaluation strategy based solely on individualized goals? *Community Mental Health Journal, 14*, 300–308.

Chan, E. K. H., Munro, D., Huang, A. H. S., Zumbo, B. D., Vojdanijahromi, R., & Ark, N. (2014). Chapter 5. Validation Practices in Counseling: Major Journals, Mattering Instruments, and the Kuder Occupational Interest Survey (KOIS). In *Validity and Valdiation in Social, Behavioral, and Health Sciences* (pp. 67–87).

Chen, M. Y., & Zumbo, B. D. (2017). Chapter 4. Ecological framework of item responding as validity evidence: An application of multilevel DIF modeling using PISA data. In *Understanding and Investigating Response Processes in Validation* (1st ed., pp. 53–69). Switzerland: Springer International Publishing.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of Validity Evidence for Educational and Psychological Tests. *Educational and Psychological Measurement, 68*, 397–412. https://doi.org/10.1177/0013164407310130

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin*, *74*, 68–80. https://doi.org/10.1037/h0029382

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302. https://doi.org/10.1037/h0040957

Cusick, A., McIntyre, S., Novak, I., Lannin, N., & Lowe, K. (2006). A comparison of goal attainment scaling and the Canadian occupational performance measure for paediatric rehabilitation research. *Developmental Neurorehabilitation*, *9*, 149–157. https://doi.org/10.1080/13638490500235581

Cytrynbaum, S., Ginath, Y., Birdwell, J., & Brandt, L. (1979). Goal attainment scaling: A critical review. *Evaluation Quarterly*, *3*, 5–40. https://doi.org/10.1177/0193841X7900300102

de Beurs, E., Lange, A., Blonk, R. W. B., Koele, P., van Balkom, A. J. L. M., & Van Dyck, R. (1993). Goal attainment scaling: An idiosyncratic method to assess treatment effectiveness in agoraphobia. *Journal of Psychopathology and Behavioral Assessment*, *15*, 357–373. https://doi.org/10.1007/BF00965038

Domene, J. F., Valach, L., & Young, R. A. (2015). Chapter 9. Action in counselling: A contextual action theory perspective. In R. Yount, J. F. Domene, & L. Valack (Eds.), *Counseling and action: Toward life-enhancing work, relationships, and identity* (pp. 152–166). New York: Springer.

Donnelly, C., & Carswell, A. (2002). Individualized outcome measures: A review of the literature. *Canadian Journal of Occupational Therapy*, *69*, 84–94. https://doi.org/10.1177/000841740206900204

Doran, G. T. (1981). There's a S.M.A.R.T. way to write management's goals and objectives. *Management Review*, *70*, 35. https://doi.org/10.1177/004057368303900411

Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, *37*, 830–837.

Elliot, A. J., & Fryer, J. W. (2008). Chapter 15. The Goal Construct in Psychology. In J. Y. Shah & W. L. Gardner (Eds.), *Handbook of Motivation Science* (pp. 235–250). New York, NY: The Guilford Press.

Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing*, *62*, 107–115. https://doi.org/10.1111/j.1365-2648.2007.04569.x

Embretson, S. E. (2016). Understanding examinees' responses to items: Implications for

measurement. *Educational Measurement: Issues and Practice, 35*, 6–22. https://doi.org/10.1111/emip.12117

Fishbach, A., & Ferguson, M. J. (2007). The goal construct in social psychology. In A. W. Kruglanski & T. E. Higgens (Eds.), *Social Psychology: Handbook of Basic Principles* (pp. 334–352). New York, NY US: Guilford Press. https://doi.org/10.1007/BF02333407

Fisher, K., & Hardie, R. J. (2002). Goal attainment scaling in evaluating a multidisciplinary pain management programme. *Clinical Rehabilitation, 16*, 871–877.

Flake, J. K., & Fried, E. I. (2019). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Preprint*, (January). https://doi.org/10.31234/osf.io/hs7wm

Fox, J. D. (2003). From products to process: An ecological approach to bias detection. *International Journal of Testing*, *3*, 21–47. https://doi.org/10.1207/S15327574IJT0301_2

Fryer, J. W., & Elliot, A. J. (2007). Stability and change in achievement goals. *Journal of Educational Psychology, 99*, 700–714. https://doi.org/10.1037/0022-0663.99.4.700

Gaasterland, C. M. W., Jansen-van der Weide, M. C., Weinreich, S. S., & van der Lee, J. H. (2016). A systematic review to investigate the measurement properties of goal attainment scaling, towards use in drug trials. *BMC Medical Research Methodology*, *16*, 99. https://doi.org/10.1186/s12874-016-0205-4

Gollwitzer, P. M. (1993). Goal Achievement: The Role of Intentions. *European Review of Social Psychology*, *4*(1), 141–185. https://doi.org/10.1080/14792779343000059

Goodwin, L. D., & Leech, N. L. (2003). The meaning of validity in the New Standards for Educational and Psychological Testing: Implications for measurement courses. *Measurement and Evaluation in Counseling and Development*, *36*, 181–191.

Gordon, J. E., Powell, C., & Rockwood, K. (1999). Goal attainment scaling as a measure of

clinically important change in nursing-home patients. *Age and Ageing*, *28*, 275–281. https://doi.org/10.1093/ageing/28.3.275

Guion, R. M. (1980). On Trinitarian doctrines of validity. *Professional Psychology: Research and Practice*, *11*, 385–398. https://doi.org/10.1037/0735-7028.11.3.385

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, *7*, 238–247. https://doi.org/10.1037/1040-3590.7.3.238

Heavlin, W. D., Lee-Merrow, S. W., & Lewis, V. M. (1982). The psychometric foundations of goal attainment scaling. *Community Mental Health Journal*, *18*, 230–241. https://doi.org/10.1007/BF00754339

Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, *15*, 1277–1288. https://doi.org/10.1177/1049732305276687

Hubley, A. M. (2017). Expanding Views on Response Processes Evidence for Validity. *Measurement: Interdisciplinary Research and Perspectives*, *15*, 140–142. https://doi.org/10.1080/15366367.2017.1404366

Hubley, A. M., & Zumbo, B. D. (1996). A dialectic on validity: Where we have been and where we are going. *Journal of General Psychology*, *123*, 207–215.

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research*, *103*, 219–230.

Hubley, A. M., & Zumbo, B. D. (2013). Psychometric characteristics of assessment procedures: An overview. In *K. F. Geisinger (Ed.). APA handbook of testing and assessment in psychology* (Vol. 1, pp. 3–19). Washington, D.C: American Psychological Association Press.

Hubley, A. M., & Zumbo, B. D. (2017). Chapter 1. Response Processes in the Context of Validity: Setting the Stage. In Bruno D. Zumbo & A. M. Hubley (Eds.), *Understanding and Investigating Response Processes in Validation* (1st ed., pp. 1–13). Switzerland: Springer International Publishing.

Hurn, J., Kneebone, I., & Cropley, M. (2006). Goal setting as an outcome measure: A systematic review. *Clinical Rehabilitation*, *20*, 756–772. https://doi.org/20/9/756 [pii]\r10.1177/0269215506070793

Jones, M. C., Walley, R. M., Leech, A., Paterson, M., Common, S., & Metcalf, C. (2006). Using goal attainment scaling to evaluate a needs-led exercise programme for people with severe and profound intellectual disabilities. *Journal of Intellectual Disabilities*, *19*, 317–335.

Joyce, B. M., Rockwood, K., & Mate-Kole, C. C. (1994). Use of Goal Attainment Scaling in brain injury in rehabilitation hospital. *American Journal of Physical Medicine & Rehabilitation*, *73*, 10–14.

Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, *64*, 425–461.

Kane, M. T. (1990). An argument-based approach to validation. *Psychological Bulletin*, *112*, 527–535.

Kane, M. T. (2001). Current concerns in validity theory. *Language Learning*, *38*, 319–342.

Kane, M. T. (2006). Content-related validity evidence. In *Downing, S.M. & Haladyna, T.M. (Ed.). Handbook of test development* (pp. 131–153). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. In *In R. W. Lissitz [Ed.], The concept of validity: Revisions, new directions and applications.* (pp. 39–64).

Charlotte, NC: Information Age Publishing.

Kielhofner, G. (2008). *Model of human occupation* (4th ed.). Philadelphia: Lippincott, Williams & Wilkins.

Kiresuk, T., Smith, A., & Cardillo, J. E. (Eds.). (1994). *Goal attainment scaling: Applications, theory, and measurement*. New York, NY: Lawrence Erlbaum Associates Inc.

Kiresuk, T J, Lund, S. H., & Larsen, N. E. (1982). Measurement of goal attainment in clinical and health care programs. *Drug Intelligence & Clinical Pharmacy*, *16*, 145–153.

Kiresuk, T. J., & Sherman, R. E. (1968). Goal attainment scaling: A general method for evaluating comprehensive community mental health programs. *Community Mental Health Journal*, *4*, 443–453. https://doi.org/10.1007/BF01530764

Koller, I., Levenson, M. R., & Glück, J. (2017). What do you think you are measuring? A mixed-methods procedure for assessing the content validity of test items and theory-based scaling. *Frontiers in Psychology*, *8*, 1–20. https://doi.org/10.3389/fpsyg.2017.00126

Krasny-Pacini, A., Hiebel, J., Pauly, F., Godon, S., & Chevignard, M. (2013). Goal Attainment Scaling in rehabilitation: A literature-based update. *Annals of Physical and Rehabilitation Medicine*, *56*, 212–230. https://doi.org/10.1016/j.rehab.2013.02.002

Krasny-Pacini, A., Evans, J., Sohlberg, M. M., & Chevignard, M. (2016). Proposed criteria for appraising goal attainment scales used as outcome measures in rehabilitation research. *Archives of Physical Medicine and Rehabilitation*, *97*, 157–170. https://doi.org/10.1016/j.apmr.2015.08.424

Launeanu, M., & Hubley, A. M. (2017). Chapter 6. Some observations on response processes research and its future theoretical and methodological directions. In *Understanding and Investigating Response Processes in Validation* (1st ed., pp. 93–113). Switzerland: Springer

International Publishing.

Law, M., Baptiste, S., McColl, M., Opzoomer, A., Polatajko, H., & Pollock, N. (1990). The Canadian occupational performance measure: An outcome measure for occupational therapy. *Canadian Journal of Occupational Therapy., 57*, 82–87.

Leighton, J. P. (2013). Item difficulty and interviewer knowledge effects on the accuracy and consistency of examinee response processes in verbal reports. *Applied Measurement in Education*, *26*, 136–157. https://doi.org/10.1080/08957347.2013.765435

Leighton, J. P. (2015). Accounting for affective states in response processing data: Impact for validation. *Paper presented at the annual meeting of the National Council on Measurement in Education (NCME)*, Chicago, IL, USA

Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice*. https://doi.org/10.1111/j.1745-3992.2007.00090.x

Leighton, J. P., Tang, W., & Guo, Q. (2017). Chapter 8. Response processes and validity evidence: Controlling for emotions in think aloud interviews. In Bruno D. Zumbo & A. M. Hubley (Eds.), *Understanding and Investigating Response Processes in Validation* (1st ed., pp. 137–158). Switzerland: Springer International Publishing.

Levack, W. M., Weatherall, M., Hay-Smith, J. C., Dean, S. G., McPherson, K., & Siegert, R. J. (2016). Goal setting and strategies to enhance goal pursuit in adult rehabilitation: Summary of a Cochrane systematic review and meta-analysis. *European Journal of Physical and Rehabilitation Medicine, 52*, 400–416.

Locke, E. A. (1968). Toward a theory of task motivation and incentives. *Organizational*

Behavior and Human Performance, 3, 157–189. https://doi.org/10.1016/0030-5073(68)90004-4

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *The American Psychologist, 57*, 705–717. https://doi.org/10.1037/0003-066X.57.9.705

Locke, E. A., Shaw, K. N., Saari, L. M., & Latham, G. P. (1981). Goal setting and task performance: 1969-1980. *Psychological Bulletin, 90*, 125–152. https://doi.org/10.1037/0033-2909.90.1.125

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635–694. https://doi.org/10.2466/pr0.1957.3.3.635

Maddox, B. (2017). Talk and Gesture as Process Data. *Measurement: Interdisciplinary Research and Perspectives, 15*, 113–127. https://doi.org/10.1080/15366367.2017.1392821

Maddox, B., & Zumbo, B. D. (2017). Chapter 10 - Observing testing situations: Validation as jazz. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and Investigating Response Processes in Validation* (1st ed., pp. 179–192). Switzerland: Springer International Publishing.

Malec, J. F. (1999). Goal attainment scaling in rehabilitation. *Neuropsychological Rehabilitation, 9*, 253–275. https://doi.org/10.1080/096020199389365

Mann, T., de Ridder, D., & Fujita, K. (2013). Self-regulation of health behavior: social psychological approaches to goal setting and goal striving. *Health Psychology, 32*, 487–498.

Mannion, A. F., Caporaso, F., Pulkovski, N., & Sprott, H. (2010). Goal attainment scaling as a measure of treatment success after physiotherapy for chronic low back pain. *Rheumatology,*

49, 1734–1738. https://doi.org/10.1093/rheumatology/keq160

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York, NY: Routledge.

Marshall, S. K., Zaidman-Zait, A., Domene, J. F., & Young, R. A. (2012). Qualitative action-project method in family research. *Journal of Family Theory & Review*, *4*, 160–173. https://doi.org/10.1111/j.1756-2589.2012.00117.x

Mayring, P. (2000). Qualitative Content Analysis. *Forum: Qualitative Social Research*, *1*(2), 1–10. https://doi.org/10.1111/j.1365-2648.2007.04569.x

Mcgaghie, W. C., & Menges, R. J. (1975). Assessing Self-Directed Learning. *Teaching of Psychology*, *2*, 56–59.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, *30*, 955–956.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*, 1012–1027. https://doi.org/10.1037/0003-066X.35.11.1012

Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*, 5–11.

Messick, S. (1989b). Validity. In *R. L. Linn (Ed.), Educational measurement (3rd ed)* (pp. 13–103). New York, NY: Macmillan.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, *50*, 741–749.

Messick, S. (1994). The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher*, *23*(2), 13–23. https://doi.org/10.3102/0013189X023002013

Middel, B., & van Sonderen, E. (2002). Statistical significant change versus relevant or

important change in (quasi) experimental design: Some conceptual and methodological problems in estimating magnitude of intervention-related change in health services research. *International Journal of Integrated Care*, *2*, 1–18.

Milne, J. L., Robert, M., Tang, S., Drummond, N., & Ross, S. (2009). Goal achievement as a patient-generated outcome measure for stress urinary incontinence. *Health Expectations*, *12*, 288–300. https://doi.org/10.1111/j.1369-7625.2009.00536.x

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Grp, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, *6*. https://doi.org/10.1371/journal.pmed.1000097

Newton, P. E., & Shaw, S. D. (2013). Standards for talking and thinking about validity. *Psychological Methods*, *18*, 301–319. https://doi.org/10.1037/a0032969

O'Leary, T. M., Hattie, J. A. C., & Griffin, P. (2017). Actual interpretations and use of scores as aspects of validity. *Educational Measurement: Issues and Practice*, *36*, 16–23. https://doi.org/10.1111/emip.12141

Padilla, J.-L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, *26*, 136–144. https://doi.org/10.7334/psicothema2013.259

Padilla, J.-L., & Leighton, J. P. (2017). Chapter 12. Cognitive interviewing and think aloud methods. In B. D Zumbo & A. M. Hubley (Eds.), *Understanding and Investigating Response Processes in Validation* (1st ed., pp. 211-228). Switzerland: Springer International Publishing.

Palisano, R. J., & Gowland, C. (1993). Validity of goal attainment scaling in infants with motor delays. *Physical Therapy*, *10*, 651.

Palisano, R. J., Haley, S. M., & Brown, D. A. (1992). Goal Attainment Scaling as a measure of

change in infants with motor delays. *Physical Therapy*, *72*, 432–437.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Rock, B. D. (1987). Goal and outcome in social work practice. *Social Work*, *32*, 393–398. https://doi.org/10.1093/sw/32.5.393

Rockwood, K., Stolee, P., Howard, K., & Mallery, L. (1996). Use of Goal Attainment Scaling to measure treatment effects in an anti-dimentia drug trial. *Neuroepidemiology*, *15*, 330–338.

Rockwood, K. (1994). Setting goals in geriatric rehabilitation and measuring their attainment. *Reviews in Clinical Gerontology*, *4*, 141–149. https://doi.org/10.1017/S0959259800003737

Rushton, P. W., & Miller, W. C. (2002). Goal attainment scaling in the rehabilitation of patients with lower-extremity amputations: A pilot study. *Archives of Physical Medicine & Rehabilitation*, *83*, 771–775.

Saavedra, R., Earley, P. C., & Van Dyne, L. (1993). Complex interdependence in task-performing groups. *Journal of Applied Psychology*, *78*, 61–72. https://doi.org/10.1037/0021-9010.78.1.61

Sakzewski, L., Boyd, R., & Ziviani, J. (2007). Clinimetric properties of participation measures for 5 to 13 year old children with cerebral palsy: A systematic review. *Developmental Medicine & Child Neurology*, *49*, 232–240.

Schlosser, R. W. (2004). Goal attainment scaling as a clinical measurement technique in communication disorders: A critical review. *Journal of Communication Disorders*, *37*, 217–239. https://doi.org/10.1016/j.jcomdis.2003.09.003

Schuwirth, L. (2014). From structured, standardized assessment to unstructured assessment in the workplace. *Journal of Graduate Medical Education*, *6*, 165–166.

https://doi.org/10.4300/JGME-D-13-00416.1

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*, 93–105.

Schwarzer, R. (1992). Self-efficacy in the adoption and maintenance of health behaviors: Theoretical approaches and a new model. In *Self-efficacy: Thought control of action* (pp. 217–243). London: Hemisphere.

Scobbie, L., Dixon, D., & Wyke, S. (2011). Goal setting and action planning in the rehabilitation setting: Development of a theoretically informed practice framework. *Clinical Rehabilitation, 25*, 468–482. https://doi.org/10.1177/0269215510389198

Shankar, S., Marshall, S. K., & Zumbo, B. D. (*in press*). A systematic review of validation practices for the Goal Attainment Scaling measure. *Journal of Psychoeducational Assessment*. https://doi.org//10.1177/0734282919840948

Shankar, S., Miller, W. C., Roberson, N. D., & Hubley, A. M. (2019) Assessing patient motivation for treatment: A systematic review of available tools, their measurement properties and conceptual definition. *Journal of Nursing Measurement, 27*, 177-209. http://dx.doi.org/10.1891/1061-3749.27.2.177

Shear, B. R., & Zumbo, B. D. (2014). What counts as evidence: A review of validity studies in Educational and Psychological Measurement. In B. D. Zumbo & E. K. Chan (Eds.), *Validity and Validation in Social, Behavioral, and Health Sciences* (1st ed., pp. 91–111). Switzerland: Springer International Publishing.

Shefler, G., Canetti, L., & Wiseman, H. (2001). Psychometric properties of goal-attainment scaling in the assessment of Mann's time-limited psychotherapy. *Journal of Clinical Psychology, 57*, 971–979.

Sireci, S. G., & Sukin, T. (2013). Chapter 4. Test validity. In K. F. Geisinger (Ed.), *APA Handbook of Testing and Assessment in Psychology: Vol. 1 Test Theory and Testing Assessment in Industrial and Organizational Psychology* (pp. 61–84). American Psychological Association.

Sireci, S.G. (1998). The construct of content validity. *Social Indicators Research, 45,* 83–117. https://doi.org/10.1007/sl

Sireci, S. G. (2012). "De-constructing" test validation. *Paper presented at the annual conference of the National Council on Measurement in Education,* Vancouver, BC.

Stake, R. E. (2005). *Multiple case study analysis.* New York, NY: The Guilford Press.

Steenbeek, D., Ketelaar, M., Galama, K., & Gorter, J. W. (2007). Goal attainment scaling in paediatric rehabilitation: A critical review of the literature. *Developmental Medicine and Child Neurology, 49,* 550–556. https://doi.org/10.1111/j.1469-8749.2007.00550.x

Stolee, P., Stadnyk, K., Myers, A. M., & Rockwood, K. (1999). An individualized approach to outcome measurement in geriatric rehabilitation. *The Journals of Gerontology, 54,* M641–M647. https://doi.org/10.1093/gerona/54.12.M641

Stolee, P, Rockwood, K., Fox, R. A., & Streiner, D. L. (1992). The use of goal attainment scaling in a geriatric care setting. *J Am Geriatr Soc, 40,* 574–578. https://doi.org/10.1111/j.1532-5415.1992.tb02105.x

Stolee, P., Awad, M., Byrne, K., Deforge, R., Clements, S., & Glenny, C. (2012). A multi-site study of the feasibility and clinical utility of Goal Attainment Scaling in geriatric day hospitals. *Disability and Rehabilitation, 34,* 1716–1726. https://doi.org/10.3109/09638288.2012.660600

Sumsion, T. (2000). A revised occupational therapy definition of client-centred practice. *British*

Journal of Occupational Therapy, 63, 304–309.

Tenopyr, M. L. (1977). Content-construct confusion. *Personnel Psychology, 30,* 47–54.

Terwee, C. B., Dekker, F. W., Wiersinga, M., Prummel, M. F., & Bossuyt, P. M. (2003). On assessing responsiveness of health-related quality of life instruments: Guidelines for instrument evaluation. *Quality of Life Research, 12,* 349–362.

Thomas, D. R., & Zumbo, B. D. (2012). Difference scores from the point of view of reliability and repeated-measures ANOVA: In defense of difference scores for data analysis. *Educational and Psychological Measurement, 72,* 37–43. https://doi.org/10.1177/0013164411409929

Tourangeau, R. (1984). Cognitive aspects of survey design: Building a bridge between disciplines. In T. Jabine, M. Straf, J. Tanur, & R. Tourangeau (Eds.), *Cognitive science and survey methods: A cognitive perspective*. (pp. 73–100). Washington, D.C.: National Academy Press.

Turner-Stokes, L., Fheodoroff, K., Jacinto, J., Maisonobe, P., & Zakine, B. (2013). Upper limb international spasticity study: Rationale and protocol for a large, international, multicentre prospective cohort study investigating management and goal attainment following treatment with botulinum toxin A in real-life clinical practice. *BMJ Open, 3,* 1–12. https://doi.org/10.1136/bmjopen-2012-002230

Valach, L., & Young, R. A. (2002). Contextual action theory in career counselling: Some misunderstood issues. *Canadian Journal of Counselling, 36,* 97–112.

Valach, L., & Young, R. A. (2009). Interdisciplinarity in vocational guidance: An action theory perspective. *International Journal for Educational and Vocational Guidance, 9,* 85–99. https://doi.org/10.1007/s10775-009-9156-1

Valach, L., Young, R. A., & Domene, J. F. (2015). Chapter 10. Current counseling issues from the perspective of contextual action theory. In R. Yount, J. F. Domene, & L. Valack (Eds.), *Counseling and action: Toward life-enhancing work, relationships, and identity* (pp. 165–193). New York: Springer.

Vogt, D. S., King, D. W., & King, L. A. (2004). Focus groups in psychological assessment: Enhancing content validity by consulting members of the target population. *Psychological Assessment, 16,* 231–243.

von Cranach, M., Kalbermatten, U., Indermuhle, K., & Gugler, B. (1982). *Goal-directed action (M. Turton, Trans.)*. London: Academic Press. (Original work published 1980).

Vu, M., & Law, A. V. (2012). Goal-attainment scaling: A review and applications to pharmacy practice. *Research in Social and Administrative Pharmacy, 8,* 102–121. https://doi.org/10.1016/j.sapharm.2011.01.003

Wall, J. M., Law, A. K., Zhu, M., Munro, D., Parada, F., & Young, R. A. (2016). Understanding goal-directed action in emerging adulthood: Conceptualization and method. *Emerging Adulthood, 4,* 30–39. https://doi.org/10.1177/2167696815610695

Webb, T. L., & Sheeran, P. (2006). Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence. *Psychological Bulletin, 132,* 249–268. https://doi.org/10.1037/0033-2909.132.2.249

Webb, T. L., & Sheeran, P. (2007). How do implementation intentions promote goal attainment? A test of component processes. *Journal of Experimental Social Psychology, 43,* 295–302. https://doi.org/10.1016/j.jesp.2006.02.001

Willer, B., & Miller, G. (1976). On the validity of goal attainment scaling as an outcome measure in mental health. *Public Health Briefs, 66,* 1197–1198.

Woodward, C. A., Santa-Barbara, J., Levin, S., & Epstein, N. B. (1978). The role of goal attainment scaling in evaluating family therapy outcome. *American Journal of Orthopsychiatry, 48*, 41–49.

Yip, A. M., Gorman, M. C., Stadnyk, K., Mills, W. G. M., Macpherson, K. M., & Rockwood, K. (1998). A standardized menu for goal attainment scaling in the care of frail elders. *The Gerontologist, 38*, 735–742.

Young, R. A., Antal, S., Bassett, M. E., Post, A., Devries, N., & Valach, L. (1999). The joint actions of adolescents in peer conversations about career. *Journal of Adolescence, 22*, 527–538. https://doi.org/10.1006/jado.1999.0246

Young, R. A., Marshall, S. K., Stainton, T., Wall, J. M., Curle, D., Zhu, M., … Zaidman-Zait, A. (2018). The transition to adulthood of young adults with IDD: Parents' joint projects. *Journal of Applied Research in Intellectual Disabilities, 31*, 224–233. https://doi.org/10.1111/jar.12395

Young, R. A., Valach, L., Ball, J., Paseluikho, M. A., Wong, Y. S., DeVries, R. J., … Turkel, H. (2001). Career development in adolescence as a family project. *Journal of Counseling Psychology, 48*, 190–202. https://doi.org/10.1037/0022-0167.48.2.190

Young, R. A., Valach, L., & Collin, A. (2002). A contextualist explanation of career. In D. B. and Associates (Ed.), *Career choice and development* (4th ed., pp. 206–252). San Francisco: Jossey- Bass.

Young, R. A., Valach, L., & Domene, F. (2005). The action–project method in counseling psychology. *Journal of Counseling Psychology, 52*, 215–223. https://doi.org/10.1037/0022-0167.52.2.215

Young, R. A., Marshall, S. K., & Valach, L. (2007). Cultural sensitivity, career theories, and counseling. *The Career Development Quarterly, 56*

Zumbo, B.D. (2007a). Validity: Foundational issues and statistical methodology. *In C. R. Rao & S. Sinharay (Eds.). Handbook of statistics, Vol. 26: Psychometrics*. Amsterdam, Netherlands: Elsevier Science.

Zumbo, B. D. (2007b). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*, 223–233. https://doi.org/10.1080/15434300701375832

Zumbo, B.D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In *R. W. Lissitz (Ed.), The concept of validity: Revisions, new directions and applications* (Vol. 48, pp. 65–82). Charlotte, NC: Information Age. https://doi.org/10.1111/j.1745-3984.2011.00155.x

Zumbo, B. D. (2015, November). Consequences, side effects and the ecology of testing: Keys to considering assessment "in vivo." Plenary address, the Annual Meeting of the Association for Educational Assessment – Europe (AEA- Europe), Glasgow, Scotland. Retrieved from https://youtu.be/0L6Lr2BzuSQ

Zumbo, B.D. (2017a). Trending away from routine procedures, toward an Ecologically Informed In Vivo View of Validation Practices. *Measurement: Interdisciplinary Research and Perspectives, 15*, 137–139. https://doi.org/10.1080/15366367.2017.1404367

Zumbo, B. D. (2017b). Chapter 19. On models and modeling in measurement and validation studies. In B. D Zumbo & A. M. Hubley (Eds.), *Understanding and Investigating Response Processes in Validation* (1st ed., pp. 363–370). Switzerland: Springer International Publishing.

Zumbo, B.D., & Hubley, A. M. (2017). *Understanding and Investigating Response Processes in Validation*. (B. D. Zumbo & A. M. Hubley, Eds.) (1st ed.). Switzerland: Springer International Publishing.

Zumbo, B. D., & Chan, E. K. H. (Eds.). (2014). *Validity and Validation in Social, Behavioral and Health Sciences*. Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-07794-9

Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B. R., Olvera Astivia, O. L., & Ark, T. K. (2015). A methodology for Zumbo's third generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly*, *12*, 136–151. https://doi.org/10.1080/15434303.2014.972559

Appendices

Appendix A: Goal Attainment Scaling instructions

Kiresuk et al. (1994) outline a step-by-step process to guide the development and scaling of goals, as an example of their use in a mental health setting. These steps include the following:

1. Identify the issues that will be addressed during treatment.
2. Translate the problems into at least 3 goals.
3. Choose a title for each goal to convey the intent of the goal (e.g. reduce anxiety attacks).
4. Select a "behavior, affective state, skill or process" that represents the goal and also indicates progress towards the goal (e.g. for a client that may experience depression, an indicator might include sleep disturbance).
5. Identify an expected outcome level for the goal (e.g. sleep 7.5 to 8.5 hours).
6. Review the expected outcome level of the goal to ensure it is consistent with the goal title and can be interpreted by others.
7. Specify outcome levels that indicate "somewhat more and somewhat less" than expected levels of goal achievement.
8. Specify outcome levels that indicate "much more and much less" than expected levels of goal achievement.
9. Repeat the scaling steps for each of the goals.

Appendix B: Goal Attainment Scaling Guide

LEVEL OF ATTAINMENT	Goal 1 (example) Improving Sleep Routines	Goal 2	Goal 3
-2 Much less than expected	Sleeping 6-8 hours/night 0 times per week		
-1 Somewhat less than expected	Sleeping 6-8 hours/night once per week		
0 Expected level of outcome	Sleeping 6-8 hours/night 2 times per week		
+1 Somewhat more than expected	Sleeping 6-8 hours/night 3-4 times per week		
+2 Much more than expected	Sleeping 6-8 hours/night every night of the week		
COMMENTS			

Appendix C: Defining and coding for validity evidence

Source of validity evidence	Definition	Coding
Test Content	The construct has been clearly identified and defined, and content experts were consulted.	(i) What construct was identified and if yes, what was the definition? (ii) Were content experts consulted or mentioned (yes/no). Experts were considered broadly, and may include: teachers, therapists, patients, students or family members
Response Process	Whether theory was examined or individual responses were systematically tested.	(i) Was theory used or mentioned and if yes, what was it? (ii) Were individual responses systematically tested (yes/no). If response processes were tested, how this was tested (e.g. cognitive)? If not, were interactions between individuals and the GAS measure considered?
Internal Structure	Any statistical technique to determine whether the GAS reflects the construct it proposes to measure (e.g. factor analyses)	(i) Were any statistics that tests for internal structure reported or measured? (yes/no)
Relations with other variables	Evidence for how the construct is related to other variables. Responsiveness and sensitivity to change (as a relation to its previous score) was also coded.	(i) Was this source of validity reported (yes/no) and if so, what was it called? This was coded as: convergent, divergent, criterion-predictive, criterion-concurrent, criterion-group differences, generalizations, discriminant, nomological network, construct validity, other, unsure/not clear. (ii) Was resposiveness or or sensitivity to change reported? (yes/no)
Consequences	Included positive or negative consequences of GAS. Evidence that pertained to how the score was interpreted or other evidence of the score's applied purpose and utility this was noted (Messick, 1995).	(i) Were consequences reported? (yes/no) (ii) What evidence was provided for score's applied purpose (e.g. as a change score)?

Appendix D: Case study for client-actor to role play

Client (Mr. Smith) is a 39-yr old male, who was referred to an Occupational Therapist by the psychiatrist at a community mental health clinic in Vancouver. He was referred to an Occupational Therapist to explore rehabilitation and goals. The client has a history of depression following trauma from a past incident. Six years ago, the client was admitted to hospital for suicide ideation and attempt but has been seeing the psychiatrist in the community clinic for the past several years. The client is married with two young children (6 and 8 years old) and has explained that he is having some difficulty with managing parenting of two young kids. The client is currently unemployed and is a smoker. The client explained at intake that he has some future aspirations and is possibly interested in going back to school or finding a job.