A FRAMEWORK FOR VALIDATION OF THE USE OF PERFORMANCE ASSESSMENT IN SCIENCE

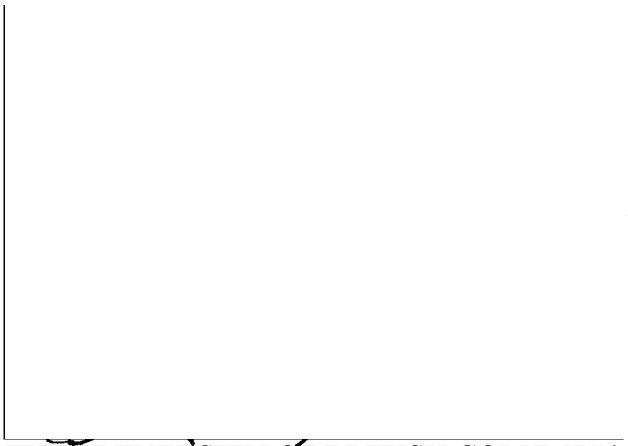
by

ANTHONY WILLIAM BARTLEY

B.A. The University of Essex
M.Sc. The University of Kent at Canterbury

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
in
THE FACULTY OF GRADUATE STUDIES
(Centre for the Study of Curriculum and Instruction)

We accept this thesis as conforming
to the required standard



THE UNIVERSITY OF BRITISH COLUMBIA

April, 1995

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

(Signature) _____

Centre for the Study of Curriculum and Instruction
~~Department of~~ _____

The University of British Columbia
Vancouver, Canada

Date 26 April 1995

DE-6 (2/88)

Abstract

The assessment of learning in school science is important to the students, educators, policy makers, and the general public. Changes in curriculum and instruction in science have led to greater emphasis upon alternative modes of assessment. Most significant of these newer approaches is "performance assessment", where students manipulate materials in experimental situations. Only recently has the development of performance assessment procedures, and the appropriate strategies for interpreting their results, received substantial research attention.

In this study, educational measurement and science education perspectives are synthesized into an integrated analysis of the validity of procedures, inferences and consequences arising from the use of performance assessment. The Student Performance Component of the 1991 B.C. Science Assessment is offered as an example. A framework for the design, implementation, and interpretation of hands-on assessment in school science is presented, with validity and feasibility considered at every stage. Particular attention is given to a discussion of the influence of construct labels upon assessment design. A model for the description of performance assessment tasks is proposed. This model has the advantage of including both the science content and the science skill demands for each task. The model is then expanded to show how simultaneous representation of multiple tasks enhances the ability to ensure adequate sampling from appropriate content domains.

The main conclusion of this validation inquiry is that every aspect of performance assessment in science is influenced by the perspective towards learning in science that permeates the assessment, and that this influence must be considered at all times. Recommendations are made for those carrying out practical assessments, as well as suggestions of areas that invite further research.

Table of Contents

List of Tables

List of Figures

Acknowledgements

My thanks and appreciation goes to my wife, Jan MacPhail, her parents Donald and Emma MacPhail, and my mother Hilda Bartley who has watched and listened from afar.

My work on the Student Performance Component of the 1991 B.C. Science Assessment with Gaalen Erickson, Bob Carlisle, Karen Meyer, Lorna Blake and Ruth Stavy was a vital part of my growth. The Ministry of Education of the Province of British Columbia funded the 1991 B.C. Science Assessment; Jim Gaskill and Amy Bryden valued my work on the provincial assessment project, which led to my involvement in the workshops for district performance assessment around the province.

My committee—Gaalen Erickson, Dave Bateson and Nand Kishor—were most encouraging. I value my contact with each and all of them, particularly as I was framing the focus of the study.

The community of graduate students at U.B.C. in Math and Science (now Curriculum Studies) provided a supportive environment. Foremost among these supporters were Tony Clarke and Renee Fountain who were there at the beginning and the finale. I wish both of them well in their own careers.

CHAPTER 1 — INTRODUCTION

DESCRIPTION OF THE PROBLEM

A powerful message about the systemic nature of educational reform was sent through North America when the American Association for the Advancement of Science (AAAS) organized the 1990 Forum for School Science, *Assessment in the Service of Instruction* (Champagne, Lovitts, & Callinger, 1990). The crux of this message was that for change in curriculum and instruction in science to be effective there must also be change in assessment practices. Good assessment in science must expand from mere indicators of performance in the form of multiple-choice tests to include direct measures of performance (Lovitts and Champagne, 1990). In this dissertation I address some of the problems related to this shift in emphasis to performance-based assessment, and propose some solutions.

The history and changes of perspective in the development of large-scale multiple-choice tests are outlined by Cole (1991). At their inception it was assumed that such tests were "neutral indicators of student progress largely isolated from classroom concerns" (Cole, 1991, pp. 97-98). However, when competency in the so-called "basic skills" became an issue in the 1970's, and criterion-referenced testing programs were established with district and school scores published, these neutral indicators became an integral part of the system. Teachers perceived a need to focus on test preparation by spending more time on topics covered by the tests and felt obliged to prepare students for the tests by drilling specific item formats. Lovitts and Champagne (1990) argue that such large-scale, multiple-choice assessments serve the needs of policy-makers rather than those of classroom teachers or their students in part because assessments of this type produce aggregated data amenable to statistical manipulation. This enables politicians and senior administrators to monitor the status of science programs within a district, state, or even nation, as in the case of the National Assessment of Educational Progress (NAEP). Lovitts and Champagne

contend that assessment by multiple-choice tests emphasizes the less important aspects of scientific literacy such as recognition and recall of factual information, rather than the more significant aspects of science such as generating and testing hypotheses, designing and conducting experiments, and solving multi-step problems. Their opinion is supported in the Carnegie Commission report *In the National Interest: The Federal Government in the Reform of K--12 Math and Science Education* (1991) which identifies the emphasis upon standardized testing as leaving "students without the capacity to think quantitatively and solve problems for themselves" (p. 23). *In the National Interest* followed a series of national reports describing problems in the American school system[1], each with an emphasis upon the position of science and mathematics (Tucker, 1991) and describing the need for systemic change.

In British Columbia, a Royal Commission embarked on an extensive examination of the education system and published the report, *A Legacy for Learners: The Report of the Royal Commission on Education 1988* (Sullivan, 1988). This report led to a response from the Ministry of Education in the form of the document *Year 2000: A Framework for Learning* (Ministry of Education, 1990). *Year 2000* sets out the framework for education reform in the Province for the decade leading to the year 2000. A vital part of this proposed change is a learner-focused framework for curriculum and assessment which is defined as:

> *developmentally appropriate*, allows for *continuous learning*, provides for *self direction*, meets the *individual learning* needs of the students as much as possible and deals with matters of *relevance* to the learners. (Emphasis in original. Ministry of Education, 1990, p. 9)

---

1 *A Nation at Risk* (National Commission on Excellence in Education, 1983), *A Nation Prepared* (Carnegie Forum on Education and the Economy, 1986), *A Time For Results* (National Governors' Association, 1986), *National Goals for Education* (U.S. Department of Education, 1990) and *America's Choice: High Skills or Low Wages!* (National Center on Education and the Economy, 1991)

The 1991 B.C. Science Assessment was the first major project of the Assessment, Examinations, and Reporting Branch of the Ministry of Education to be conceived and conducted following the publication of the *Year 2000* document. The design of the Assessment reflects this new policy, particularly in the range of the components.

British Columbia has a long history of science assessments as part of the Provincial Learning Assessment Program (Hobbs, Boldt, Erickson, Quelch, & Sieben, 1980; Taylor, Hunt, Sheppy, & Stronck, 1982; Bateson, Anderson, Dale, McConnell, & Rutherford, 1986). The 1991 Science Assessment (Bateson, Erickson, Gaskell, & Wideen, 1992) with its principal focus upon Grade 4, 7 and 10 students, built upon this history. The mode of assessment was extended from primarily the use of multiple-choice questions to include open-ended written questions, a classroom observation component, a socio-scientific issues component, and a performance assessment component. These additional components enabled collection of a wide range of valuable data about science education in British Columbia. The performance component is the most significant development in the current debate about the quality of alternative assessments (Rothman, 1990a; Linn, Baker, & Dunbar, 1991), particularly with respect to performance assessment in science. For example, students from B.C. took part in the performance option of the 1991 International Assessment of Educational Progress (IAEP) in mathematics and science (Semple, 1992). The United States did not — because "the performance-based items were not of the same quality as the multiple-choice questions" (Rothman, 1990b, p. 10). Responding to this criticism, and the absence of the U.S.A. from the performance option, Rothman reports that Lapointe, the project director at the Educational Testing Service (ETS), acknowledged that there had been little research on the validity and reliability of performance assessment use. This is particularly remarkable because the ETS had coordinated the assessment.

My intent is to address the concerns regarding validity and reliability of performance assessments, in the curriculum area of science, by proposing a framework for validation. The description of student performance in science continues to evoke debate

about the use of the "processes of science" as a scaffold about which the assessment tasks are designed and student performance reported (Donnelly and Gott, 1985; Bryce and Robertson, 1985; Millar and Driver, 1987; Johnson, 1989; Millar, 1991), and the model of science which is portrayed by such an approach (Woolnough, 1989). Recent developments in the theory of validity (Messick, 1989a, 1989b, 1994; Moss, 1992, 1994; Shepard, 1993) have identified validity as a unitary concept, based upon "construct validity". The most significant development in this reconceptualization of validity is the inclusion of "consequential validity" as a component of construct validity. Consideration of consequences entails an examination of the values that give rise to the construct labels used in defining an assessment.

The Student Performance Component of the 1991 B.C. Science Assessment is used to exemplify arguments, procedures, and conclusions presented in this study. In the *British Columbia Assessment of Science 1991 Technical Report II: Student Performance Component* (Erickson, Bartley, Blake, Carlisle, Meyer, & Stavy, 1992), performance assessment has been defined as hands-on assessment of student abilities in science, using tasks that are classified as either "stations" or open-ended "investigations". "Stations" are short tasks, each focusing upon different aspects of school science: every student completes six different stations. The data for stations consist of the students' written responses. "Investigations" are more complex tasks in which students are able to use any combination of provided equipment to find a solution to a preset problem. The data for investigations consist of observers' records, students' written responses and interview responses.

HISTORICAL AND THEORETICAL PERSPECTIVES FOR THIS STUDY

My perspective for this validation inquiry is derived from current conceptions of validity (Messick, 1989a, 1989b, 1994; Moss, 1992, 1994; Shepard, 1993). Score interpretation along with consequential aspects of the assessment process make up the

significant elements of this inquiry. Hein (1990) warns science educators of the problems of abdicating responsibility or involvement in science assessment:

> Science assessment is too important to be left in the hands of only psychometricians and other test developers. As a problem that combines theoretical and real-world issues, it requires input from groups representing many perspectives. Only then will we come to solutions that have both theoretical validity and application in the real world of schools. (p. 279)

Specific emphasis is placed upon the construct labels used to describe student performance with materials in science (Bryce and Robertson, 1985; Black, 1986; Millar and Driver, 1987; Millar, 1989, 1991; Hodson, 1986). Concerns about quality and technical adequacy of performance assessments focus upon the relationship between the reliability of the data collected and the validity of the interpretations of those data (Linn et al. 1991). Reliability is seen as a necessary but not sufficient condition in the validity of data interpretation. Feldt and Brennan (1989) in their chapter entitled "Reliability" in the handbook *Educational Measurement* (Linn, 1989) warn against exaggerated concerns for reliability. They acknowledge:

> the primacy of validity in the evaluation of the adequacy of an educational measure. No body of reliability data, regardless of the methods used to analyze it, is worth very much if the measure to which it applies is irrelevant or redundant. (p. 143)

Conceptions of validity (Moss, 1992; Shepard, 1993) have shifted with changing perspectives in the philosophy of science (Messick, 1989b). While these three authors set out some primary considerations, the perspective for this validation inquiry synthesizes the work of:

1) Fredericksen and Collins (1989), who define criteria for enhancing systemic validity;
2) Linn, Baker, and Dunbar (1991) who set out expanded criteria for validity inquiry for alternative assessments;
3) The Quantitative Understanding: Amplifying Student Achievement and Reasoning (QUASAR) Project, a mathematics curriculum reform project intended to promote the acquisition of thinking and reasoning skills in mathematics. This project made

extensive use of performance assessment; the paper *Principles for Developing Performance Assessments* (Lane, Parke, & Moskal, 1992) presents its perspective.

4) Berlak, Newmann, Adams, Archbald, Burgess, Raven, and Romberg (1992), who argue for a "New Science of Educational Testing and Assessment". This group includes Archbald and Newman (1988) who introduced the phrase "authentic academic achievement" which has had a significant impact upon the measurement community; and

5) Herman (1992), who believes that "good assessment is built upon current theories of learning and cognition". She cites several authors, including Wittrock (1991), who consider meaningful learning to be "reflective, constructive, and self-regulated" (p. 75). Herman argues that student motivation is a significant factor in assessment acknowledging the importance of a "disposition to use the skills and strategies as well as the knowledge of when to apply them" (p. 75).

Validation inquiry studies for large scale alternative forms of assessment have been made in the curriculum areas of written composition (Welch, 1993) and mathematics (Magone, Cai, Silver, & Wang, 1992). Shavelson and his associates have made some progress in clarifying certain issues in the validation of alternative assessment in science (Pine, 1990; Baxter, Shavelson, Goldman, & Pine, 1992; Shavelson, Baxter, & Pine, 1992; Shavelson and Baxter, 1992; Shavelson, Baxter, & Gao, 1993; Shavelson, Gao, & Baxter, 1994). Aspects relating to scoring procedures, task stability, and sampling variability have been addressed through generalizability theory. Although much of the Shavelson group's work has been set in a research context and used a small number of tasks, the 1993 paper addresses the effects of sampling variability using some of the extensive data amassed from the 1990 California Assessment Program (CAP).

Pine (1990) examines the tension between judgements based upon assessing students' science knowledge and competence from work done over a short period of time in an assessment situation, and those judgements based upon longer term assessments, for

example those by a classroom teacher. Pine describes the Assessment of Performance Unit (APU) as "most likely the world's most experienced group in developing test questions for assessing science process skills", but he points out that "the APU does not have data to establish the validity of its questions"(1990, p. 91). Pine considers that expert judgement could be considered as "prejudice" and that "validating methodology is an area in which informed opinion will not suffice" (1990, p. 91). Moss believes that "we need to consider yet another expansion in our delimitation of the concept of validity" (1992, p. 252). This is particularly so in the curriculum area of science, where developments in assessment procedures have moved ahead of our capacity to give meaning to the data produced, i.e. the validation procedures.

Of the issues raised in the assessment literature, the validation of hands-on performance assessments in science is the one that I address in this dissertation. Linn, Baker, and Dunbar (1991) envisage that a general set of criteria used to judge the adequacy of assessments should include consequences, fairness, transfer and generalizability, cognitive complexity, content quality, content coverage, meaningfulness, and cost and efficiency. These criteria are evaluated, clarified, and expanded in considering the validation of performance assessment in science.

RESEARCH QUESTIONS

My work in this dissertation is to propose and evaluate procedures for validation in the assessment of students' hands-on performance in science. The range of evidence considered in such an approach necessarily encompasses both analytical and empirical domains. Procedures developed for and data produced in the Student Performance Component of the 1991 B.C. Science Assessment (Erickson et al., 1992) are used to exemplify the model proposed. Specifically, this validation inquiry addresses four major questions:

1. What are the essential components of a systematic framework for the development and administration of performance assessments in science?

This question should be considered a foundation question which sets out a myriad of concerns about principles and procedures that must be in place to enable valid inferences about student performance in science.

2. What are the essential characteristics of descriptors of student performance on performance tasks in science?

Question 2 addresses the issue of choice of construct labels. It requires a discussion of the identification and labelling of skills/processes and content in science, and the reasons why this approach is attractive in assessment. This is an issue that permeates the dissertation since it is at the core of validity for performance assessment in science.

3. What are the implications of using different strategies for scoring student achievement upon the interpretation of student performance?

This question expands the analysis of holistic and analytical (atomistic) scoring into an examination of construct validity, reliability, and messages about learning in science given by the use of a specific scoring system. This aspect of consequential validity is receiving greater prominence (Baxter, Shavelson, Herman, Brown, & Valadez, 1993) though little has yet been published in the field of science assessment.

4. How could/should test scores[2] in performance assessments be interpreted and used?

Question 4 is not only a "so what?" question, but also a "who?" question. The interpretation of student performance, and the consequences of that interpretation, are the crux of the matter. It is essential to discuss who should interpret student performance and the implications of their involvement. Students, teachers, administrators, parents,

---

2 The term *test score* is used here generically in the same way that Messick (1989b, p. 14) uses *score* to mean "any observed consistency" and *test* to include questionnaire, observation procedure or other assessment device. He qualifies this general usage to include qualitative and quantitative summaries not only of persons but of judgemental consistencies and attributes.

politicians, and professional evaluators all have perspectives towards the interpretation of performance. The nature of the interpretation, and the power to effect change are issues that must be considered.

SIGNIFICANCE OF THE STUDY

Performance assessment is still in its infancy in North America. Herman warns that "what we know about them [performance assessments] is relatively small compared to what we have yet to discover" (1992, p. 74). Baxter, Shavelson, Goldman, and Pine (1992) provide evidence that hands-on performance assessment appears to be measuring something different from pencil-and-paper modes of assessment. However, these authors concentrate upon procedural and technical issues rather than attempting to explain the conceptual problems in terms of a learning theory or a philosophy of science. In addressing conceptual as well as technical issues pertaining to performance assessment in science, this study will make a significant contribution in an area of expanding interest and concern.

DELIMITATIONS OF THE STUDY

This validation inquiry is expressly focused upon assessment of hands-on student performance in the context of school science. Questions 1 and 2 are set explicitly within the domain of science and are intended to extract some operational understanding, in this context, for the terms "reliability" and "construct validity". Questions 3 and 4 direct an analysis of consequential validity, focusing upon the consequences of scoring strategies and consequences of interpretation. Many of the procedures in the development and use of performance assessment tasks in mathematics, particularly in the use of manipulatives, are similar to those in science. Because of these parallels it is likely that many of the proposals for the validation of performance assessments in science are equally applicable to performance assessments in mathematics.

CHAPTER 2 — THEORIES OF VALIDITY, PROCESSES OF SCIENCE AND RELIABILITY

The theoretical rationale behind this validation inquiry of the assessment of hands-on performance in science is derived from current concepts of validity. In this chapter I discuss how the theories of validity have undergone a significant re-alignment as original "types" of validity have been absorbed into construct validity, and construct validity has been expanded to cover the consequences of test use. This expansion of validity leads to a discussion of the approaches that have been taken to re-examine how validation of performance assessments can be achieved. The breadth of the terms of reference of validity, particularly the considerations embedded in the consequences of test use, requires an analysis of the value positions implied in the choice of construct labels in science assessment. I set out the rationale behind the extensive use of "processes of science" as an organizing theme in hands-on assessment, and explore the concerns of those educators who argue against such a perspective.

Assessment data must conform to some agreed standard of reliability. For performance assessments involving the use of manipulatives, traditional measures of reliability are inappropriate. The assumptions of classical reliability theory are examined, and the application of generalizability theory is reviewed. Qualitative factors that influence reliability, including the instrumental variables arising from the use of equipment, are also considered.

VALIDITY — CHANGING CONCEPTUALIZATIONS

The theory of validity has undergone significant evolution in recent times. Of particular importance is the chapter by Messick (1989b), in the third edition of *Educational*

Measurement, which presents validity as a unitary concept. Messick defines validity as "an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment"[1] (1989b, p. 13). This is significant because Messick identifies that it is inferences and actions, together with their underlying theories, that must be validated. Messick argues that validity is a matter of degree, continuous over a range, and likely to change over time. Validity evidence gains or loses strength with new findings, and as the expected consequences of testing are realized (or not) by the actual consequences. This being the case, validation is never complete and the test developer must continue to search for evidence to make "the most reasonable case to guide both current use of the test and current research to advance understanding of what test scores mean" (Messick, 1989b, p. 13). Messick is consistent in emphasizing the need for validation studies to use multiple sources to obtain a range of evidence to support score-based inferences, as "validity is a unitary concept" (Messick, 1989b, p. 13). Messick describes validation as following the methods of science in the collection of evidence to support inferences; this is seen as hypothesis testing in the context of interpretive theories of score meaning. The major concern of validity is "to account for *consistency* in behavior, or item responses, which frequently reflects distinguishable determinants" (Emphasis in original. Messick, 1989b, p. 14). In restating this key aspect of validity, Messick reminds us that the:

> emphasis is on scores and measurements as opposed to tests or instruments because the properties that signify adequate assessment are properties of scores, not tests. Tests do not have reliabilities and validities, only test responses do. This is an important point because test responses are a function not only of the items, tasks, or stimulus conditions but of persons responding and the contexts of the measurement. (1989b, p. 14)

---

1 Evidence includes data, facts and the rationale or arguments used to justify the inferences (Messick, 1989b).

The context of the measurement is a further issue in validity research. Messick refers to earlier work by Cronbach (1971) and argues that it is the interpretation of data arising from a specified procedure that is to be validated and this validation should include an analysis of the generalizability of the interpretation, particularly over time.

A further aspect of generalizability is derived from consideration of test behaviour as a sample of a domain behaviour, or as an indicator of some other underlying process or trait[2]. This step beyond the obvious has led to the use of constructs, for example "critical thinking skills", in an attempt to describe these underlying traits. Such traits are given significant emphasis in educational measurement. Messick perceives these underlying structures as a source of much contention among measurement theorists: Behaviorists and social behaviorists interpret scores as samples of response classes[3] but trait theorists and cognitive theorists consider scores as signs of underlying processes or structures (Messick, 1989b). Whereas a trait represents a stable set of relationships, a person's state is considered likely to change across contexts and over time[4]. Test scores may be interpreted as signs of trait disposition, or internal states, or some combination of these. Whichever extreme is chosen there is a requirement for validation of the hypothesis (Messick, 1989b). In making this stipulation, Messick warns validity researchers against taking for granted a mode of interpretation and using this assumption to identify the nature of the evidence for the validation inquiry. Data do not constitute evidence. Messick (1989b, p. 16) cites Kaplan (1964, p. 375) to make this point:

---

2 "A trait is a relatively stable characteristic of a person which is consistently manifested to some degree when relevant despite considerable variation in the range of settings and circumstances" (Messick, 1989b, p. 15).

3 Messick defines response class as a "class of behaviors that reflect essentially the same changes when the person's relation to the environment is altered" (1989b, p. 15).

4 A state is considered as a temporary condition of mentality or mood, transitory level of arousal or drive (Messick, 1989b)

> What serves as evidence is the result of a process of interpretation – facts do *not* speak for themselves; nevertheless, facts must be given a hearing, or the scientific point to the process of interpretation is lost.

This perspective leads to a conceptualization of evidence as a blending of facts together with the theoretical rationale used for their interpretation. Messick compounds the issue further by adding consideration of values:

> Hence, just as data and theoretical interpretation were seen to be intimately intertwined in the concept of evidence, so data and values are intertwined in the concept of interpretation. And this applies not just to evaluative interpretation, where the roles of values is often explicit, but also to theoretical interpretations more generally, where value assumptions frequently lurk unexamined. Fact, meaning and value are thus quintessential constituents of the evidence and rationales underlying the validity of test interpretation and use. (Messick, 1989b, p. 16)

Thus all interpretations must be considered value-laden, and these values must be recognized as guiding the choice of data collection methods, i.e., the test.

Validity Evidence including Consequential Evidence

Messick (1989b, p. 16) lists six possible basic sources of validity evidence:

content of the test in relation to the content of the domain of reference;

descriptions of individual's responses;

internal structure of the test responses, i.e. relationship between items or parts of the test;

external structure of the test responses, i.e. relationship between test scores and other measures;

differences in test processes and structures over time, in different contexts or as a result of some treatment;

social consequences of interpreting test scores in a particular way with some examination of the intended outcomes, and the unintended effects.

Messick considers it important to contrast this approach to validity with the traditional categories of validity: content, criterion-related and construct. Each of these facets of validity had been given status as a **type** of validity rather than as a component of the whole. The relationship between these three facets and Messick's six sources of validity evidence is meaningful in clarifying developments in validation procedures.

Content validity is defined as "based on the professional judgements about the relevance of test content to the content of a particular behavioral domain of interest and about the representativeness with which item or task content covers that domain" (Messick, 1989b, p. 17). As considerations of content validity are not concerned with responses or scores that arise from the use of a test, Messick questions whether content validity should be described as a form of validity. He recognizes that content relevance and representation influence the nature of score inferences, but believes that validity of score interpretation must be supported by other evidence.

Criterion-related validity is concerned only with "specific test-criterion correlations" (Messick, 1989b, p. 17). Messick describes this as focusing upon a specific part or parts of a test's external structure. This may lead to a range of criterion related validities as scores are compared with other measures. For examples, see Burger and Burger (1994).

Construct validity is "based on an integration of any evidence that bears on the interpretation or meaning of the test scores" (Messick, 1989b, p. 17) The test score cannot be equated with the construct it is intended to tap, and should not be considered as defining the construct, but rather as "one of an extensible set of indicators of the construct" (Messick, 1989b, p. 17). In this sense, a construct would be "invoked as a latent variable or 'causal factor' to account for the relationships among its indicators" (Messick, 1989b, p. 17). The breadth of construct validity enables "almost any kind of information about a test" (Messick, 1989b, p. 17) to contribute to an understanding of its construct validity, but this contribution is strengthened if there is justification of the theoretical rationale behind the

interpretation of the scores. In the past, it was the pattern of relationships between internal and external test structures which made the major contribution to construct validation. Construct validity subsumes content representation and criterion-relatedness because such information contributes to score interpretation. With the consideration of the consequences of testing, the three historical "types" of validity and the evidence supporting them, have been embraced by construct validity.

Messick has developed a "progressive matrix" which illustrates the inclusion of the facets of validity, each building upon construct validity. Figure 1, presented here, is taken from the paper *Meanings and Values in Test Validation: The Science and Ethics of Assessment* (Messick, 1989a) as this schemata represents the progression with greater clarity than does the matrix in the chapter "Validity" (Messick, 1989b) in *Educational Measurement.*

	TEST INTERPRETATION	TEST USE
EVIDENTIAL BASIS	Construct Validity (CV)	CV + Relevance/Utility (R/U)
CONSEQUENTIAL BASIS	CV + Value Implications (VI)	CV + R/U + VI + Social Consequences

Figure 1. Facets of Validity as a Progressive Matrix
(Messick, 1989a)

In remarking upon the connection between the "consequences" and "functional worth" of testing, Messick presents a strong case for consideration of consequential evidence in validity studies:

> It is ironic that little attention has been paid over the years to the consequential basis of test validity, because validity has been cogently conceptualized in the past in terms of the functional worth of testing – that is, in terms of how well the test does the job it is employed to do. (Messick, 1989b, pp. 17-18)

As the influence of testing has become more pervasive, particularly in the change of role from a neutral indicator to an explicit component of educational programs, the need to

consider the consequences of testing has become more imperative. Messick considers that the consequential basis of testing is concerned with two issues, first:

> the appraisal of the value implications of the construct label, of the theory underlying test interpretation, and of the ideologies in which the theory is embedded...

and second:

> the appraisal of both potential and actual social consequences of the applied testing. (1989b, p. 20)

The first issue leads to the question of *how* to take account of values in test validation rather than *whether* values should be considered, because such action is "virtually mandatory" (Messick, 1989b, p. 58). The second issue identifies a requirement to broaden the scope of validation inquiry beyond merely looking for consequences that are immediate, positive and anticipated, but to search for effects that are long-term, negative and unanticipated, even if finding evidence of such outcomes emperils any future program of testing. It is in this latter situation that the value-system implicit in the mode of validation becomes of great significance.

Test validation is a process of inquiry (Messick, 1989b) and as such can be conceptualized in terms of some combination of the five systems of inquiry that are described by Churchman (1971) and his co-workers Mitroff and Sagasti (1973). In drawing together these five systems Churchman and his colleagues reconstruct "some of the major theories of epistemology in a way that makes them more pertinent to the knowledge-acquisition and model-building goals of the practicing scientist" (Messick, 1989b, p. 31). Churchman (1971) named these five systems after the philosophers whose work is best captured by the system of inquiry. The labels are Leibnizian, Lockean, Kantian, Hegelian, and Singerian (Messick, 1989b).

Messick adjudges the Leibnizian and Lockean approaches to be best suited for tackling well-structured problems, with the Lockean system appearing to capture the essence of validation of multiple-choice tests:

> A Lockean inquiring system entails an empirical and inductive approach to problem representation. Starting with an elementary data set of experiential judgments or observations and some consensually accepted methods for generating factual propositions inductively, an expanding network of facts is produced. The standard of validity in a Lockean inquiry system is the consensus of experts on the objectivity, or lack of bias, with respect to data and methods. (Messick, 1989b, p. 31)

Where there is no consensually agreed definition of the problem, Messick posits that the Kantian, Hegelian, or Singerian systems offer the most promise. A Kantian mode of inquiry starts with:

> at least two alternative theories or problem representations, from each of which are developed corresponding alternative data sets or fact networks. (Messick, 1989b, p. 31)

In the Kantian mode of inquiry these alternative theories may be divergent, but not antagonistic. Alternative theories may lead to multiple interpretations of the data, each with its own strengths and weaknesses. The test for validity in the Kantian system is the "goodness of fit or match between the theory and its associated data" (Messick, 1989b, p. 31). An Hegelian approach to inquiry is developed through the formulation of conflicting perspectives on the problem:

> Specifically, the Hegelian approach starts with at least two antithetical or contrary theories, which are then applied to a common data set. It is hoped that this dialectical confrontation between rival interpretations of the same data will expose the contrary assumptions of the competing models to open examination and policy debate. The guarantor of a Hegelian inquiry system is conflict. (Messick, 1989b, p. 31)

The Hegelian system of inquiry is most suitable for problems where there is little agreement upon the structure or even the nature of the problem (Messick, 1989b), particularly as it requires discussion of the underlying value assumptions. It is intended that the debate should be between ideas rather than between people, and that scientists should examine alternate perspectives to an extent that challenges their own positions. Churchman identifies the product of Hegelian inquiry as "two stories, one supporting the most prominent policy on one side, the other supporting the most prominent policy on the other

side" (1971, p. 177). A Singerian approach is one which uses multiple systems of inquiry to "observe" the target inquiry process[5]:

> A Singerian inquiring system starts with the set of other inquiring systems (Leibnizian, Lockean, Kantian, Hegelian) and applies any system recursively to another system, including itself. The intent is to elucidate the distinctive technical and value assumptions underlying each system application and to integrate the scientific and ethical assumptions of the inquiry. (Messick, 1989b, p. 32)

While Messick finds that most educational and psychological research has been either Leibnizian, Lockean or some combination of the two, he rejoices in the fertility of the "methodological heuristics" of the Singerian inquiring system (Messick, 1989b, p. 32). The similar feature of the Kantian, Hegelian and Singerian systems is the requirement of the existence of alternative theories. There must be dialogue to enable comparison between alternative or antithetical theories, and recognition that "observations and meanings are differentially theory-laden and theories are differentially value-laden" (Messick, 1989b, p. 32).

The discussion of the role of values centres around perceptions that the influence of values upon scientific inquiry lessens the worth of that inquiry. In the social sciences it has been an unambiguous expectation that researchers make explicit their own values in conducting research (Howe, 1985). To reinforce his argument that values guide approaches to validation, Messick traces the roots of the words 'valid' and 'value' back to the same Latin root, *valere*, which means "to be strong" (1989b, p. 59). The word 'value' came directly from the French verb *valoir*, meaning "to be worth", which Messick applies to the functional worth of testing.

---

5 Moss in her paper "Can There be Validity Without Reliability?" in *Educational Researcher* (March, 1994) provides an example of Singerian inquiry in discussing the hermeneutic and psychometric perspectives.

In science, and validation inquiry, the values of the investigator and to a great extent those of the specific scientific community to which the investigator belongs, are major determinants of problem selection, identification of research perspectives and theories, data collection and processing, and inferences that are made from those data. The value implications of the choice of construct labels are particularly important in terms of the "range of the implied theoretical and empirical referents" (Messick, 1989b, p. 60). Messick invokes the term *referent generality*, used earlier by Snow (1974), where there are difficulties in embracing all of the substantive features of a construct in a single or a composite measure. The use of such measures is very likely to allow hidden values extensive opportunities for pernicious growth and pervasive implications. There is tension between levels of generality: too specific at one end of the spectrum and too broad at the other.

> At one extreme is the apparent safety in using strictly descriptive labels tightly tied to behavioral exemplars in the test (such as Adding and Subtracting Two-Digit Numbers). The use of neutral labels descriptive of test tasks rather than of the processes presumably underlying task performance is a sound strategy with respect to test names to be sure (Cronbach, 1971). But with respect to construct labels, choices on this side sacrifice interpretive power and range of applicability if the construct might defensibly be viewed more broadly (e.g., as Number Facility). At the other extreme is the apparent richness of high-level inferential labels such as intelligence, creativity, or introversion. Choices on this side suffer from the mischievous value consequences of untrammeled surplus meaning.
>
> (Messick, 1989b, p. 60)

The search for a defensible balance of construct reference is seen to be governed by the range of research evidence available. However, Messick warns against carving this in stone by citing Cronbach's (1971) reminder that constructs refer to potential as well as actual relationships. Thus as new data are collected, it may be possible, or necessary, to review and modify the construct labels used to describe test behaviour, and these data must also guide the general statements about test behaviour.

As choices of construct labels are guided by theories or ideologies, these too must be considered part of the value implications. This is particularly important when one

considers the value implications that arise when alternative theories are brought to bear upon the same data. Messick argues that such effects support one of the relativists' basic points that the advocates of opposing theories appear to live in different worlds. As the boundaries between subjectivity and objectivity have become blurred, it has become more important to set out theories, ideologies[6] and world views for public scrutiny. A specific problem that Messick describes arises when two different theories share the same metaphorical perspective or model. This leads to investigators "talking past one another" by using the same language, but each interpreting according to her/his own framework.

Broader ideologies that give theories perspective and purpose must also be considered. A particular difficulty arises when there is "ideological overlay" which Messick considers to be a frequent occurrence in educational measurement and likely to cause much fallout as a result of radically different perceptions of the value implications of test interpretation and use. Even the claim of scientific behaviour must be seen as value-laden, as are the epistemic principles which guide scientific choices — predictive accuracy, internal consistence, unifying power and simplicity (Messick, 1989b, p. 62). Messick does not believe that it is necessary to go so far as to label scientific judgements as value judgements. He does however insist that there must be some empirical support or rational defence against criticism. The challenge in test validation is to excavate the value assumptions of a construct theory and its ideological foundations.

The consequential basis of test use is concerned not only with how well the test does the job it is employed to do but also the question of what other jobs the test actually does or might do, whether by intent or by accident. These consequences should be

---

6 Messick defines an ideology as "a complex configuration of shared values, affects, and beliefs that provides, among other things, an existential framework for interpreting the world – a 'stage-setting', as it were for viewing the human drama in ethical, scientific, economic or whatever terms" (1989b, p. 62).

considered at all levels – individual, institutional, and systemic or societal. Messick states the point eloquently:

> Although appraisal of the intended ends of testing is a matter of social policy, it is not only a matter of policy formulation but also of policy evaluation that weighs all of the outcomes and side effects of policy implementation by means of test scores. Such evaluation of the consequences and side effects of testing is a key aspect of the validation of test use. (1989b, p. 85)

For example, unintended effects may appear as gender or ethnic differences in score distributions. In the case of a test to be used for selection screening, such differences have some impact upon the functional worth of the selection process. Messick (1989b, p. 85) considers that it is important to distinguish between issues of test *invalidity* and test *validity* in this discussion[7]. Test invalidity arises when irrelevant sources of test and criterion variance overpower other relevant properties of the instrument. When gender or ethnic-related score differences arise because of valid properties of the construct tapped, then such differences contribute to score meaning and are a reflection of test validity. The challenge for validation is the identification of the irrelevant causes of variance, and consideration of the consequences of such identification upon interpretation and actions. Many of the factors that are seen to cause variance must be discounted because of political or social policies; these lead to special pressure upon professional judgement in guiding the use and interpretation of tests. Messick captures the issue:

> The point is that the functional worth of the testing depends not only on the degree to which the intended purposes are served but also on the consequences of the outcomes produced, because the values captured by the outcomes are at least as important as the values unleashed by the goals. (1989b, p. 85)

Messick perceives that testing of consequences could parallel the testing of constructs in the presentation of alternative hypotheses. In this case, counter proposals would be used to

---

7 I am using the term 'test validity' in the manner set out by Messick (1989b) to focus upon the interpretation of scores and the validity of test use.

expose "key technical and value assumptions of the original proposal to critical evaluation" (Messick, 1989b, p. 86). These counter proposals could involve a range of alternative assessment techniques which tap the construct through different interpretive methodologies, and could also include an analysis of the social consequences of not testing at all! In this respect Messick considers that the recognition of alternative perspectives about standards and about social values to be served by assessment methods is laudable. Such recognition would enable, even encourage, investigators to use multiple modes of inquiry to search for a range of effects.

Validation Procedures for Performance Assessments

The central dilemma in validating performance assessments is identified by Moss:

> Performance assessments present a number of validity problems not easily handled with traditional approaches and criteria for validity research. The assessments typically permit students substantial latitude in interpreting, responding to, and perhaps designing tasks; they result in fewer independent responses, each of which is complex, reflecting integration of multiple skills and knowledge; and they require expert judgement for evaluation. Consequently, meeting criteria related to such validity issues as reliability, generalizability, and comparability of assessments – at least as they are typically defined and operationalized becomes problematic. This results in a tension between traditionally accepted criteria for validity and criteria that derive from concerns about the instructional consequences of assessment, such as "authenticity" (Newmann, 1990), "directness" (Fredericksen and Collins, 1989), or "cognitive complexity" (Linn, et al., 1991), which are commonly invoked when arguments for the value of performance assessments are made. (1992, p. 230)

Moss argues that developments in the philosophy of validity, particularly the inclusion of consequential aspects, provide some theoretical support for performance assessment. The problem that Moss identifies is to find "the appropriate set of criteria and standards to simultaneously support the validity of an assessment-based interpretation and the validity of its impact upon the educational system" (1992, p. 230). In her examination of the *Standards for Educational and Psychological Testing* [American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 1985] Moss finds that they "stop short of articulating

the centrality of construct validity and the importance of considering the social consequences of any validity effort" (1992, p. 232). Recognizing that the *Standards* is a political document produced by a group that represents a large and varied constituency, Moss cites Messick's lament for the missed opportunity to move the measurement field forward in the consideration of construct validity as a unifying force, and the appraisal of value implications and social consequences. The argument that the "consequences of performance assessments are likely to be more beneficial to teaching and learning than the consequences of multiple-choice assessment used alone" is identified by Moss (1992, p. 248) as a unifying theme in the work of Fredericksen and Collins (1989) and Linn, Baker, and Dunbar (1991). Both these sets of authors perceive that

> performance assessments will not fare as well as multiple-choice assessment in terms of traditional criteria, particularly those that emphasize comparability, internal consistency of scores across items and readings and efficiency. (Moss, 1992, p. 249)

The positions and proposals of these authors is presented below.

Frederiksen and Collins discuss the effects of assessment practices upon the educational system in their paper, *A Systems Approach to Educational Testing* (1989). They address the issue of the validity of educational tests in an educational system in which approaches to curriculum and instruction, and student learning strategies are modified to produce higher scores. Frederiksen and Collins question the validity of introducing tests into such a dynamic system which adapts itself to the characteristics of the test. They perceive that a challenge to validity is associated with:

> the instructional changes engendered by the use of the test and whether or not they contribute to the development of knowledge and/or skills that the test purportedly measures. (1989, p. 27)

Frederiksen and Collins introduce the term *systemic validity* of a test to extend the notion of construct validity to allow for consideration of the effects of instructional change brought about by the introduction of a test into an educational system. Frederiksen and Collins state that:

> A systemically valid test is one that introduces in the educational system curricular and instructional changes that foster the development of cognitive skills that the test is designed to measure. Evidence for systemic validity would be an improvement in those skills after the test has been in place within the educational system for a period of time. (1989, p. 27)

The conditions for systemic validity pose particular problems in the construction of tests, most specifically in accounting for the evolution of instructional strategies and student learning engendered by the use of the test. Frederiksen and Collins believe that attempts to reduce or slow these modifications to instruction and learning are bound to fail, particularly because such an approach would "deny the educational system the ability to capitalize on one of its greatest strengths: to invent, modify, assimilate, and in other ways improve instruction as a result of experience" (1989, p. 28). They conclude that the most appropriate approach is to recognize that various components of the system will react by developing tests that "*directly reflect and support the development of the aptitudes and traits they are supposed to measure*" (Italics by authors, Frederiksen and Collins, 1989, p. 28).

Frederiksen and Collins identify two characteristics of tests that they believe will facilitate educational improvement. These are (1) directness of the cognitive skill of interest in the performance of some extended task and (2) the degree of subjectivity or judgement, analysis and reflection required on the part of the person who assigns the score. Direct assessment, argue Frederiksen and Collins, is systemically valid because "instruction that improves the test score will also have improved performance on the extended task and the expression of the cognitive skill within the task context" (p. 29). The argument in favour of subjective scoring has the prerequisite that scorers understand how to use the scoring categories. This has the prior requirement that scorers must be taught how to use the categories consistently. This condition, Frederiksen and Collins maintain, will lead to the development of training materials that will become a medium for communication of the critical traits that are to be promoted through the use of the assessment. They cite the example of the writing tasks developed for the National Assessment of Educational

Progress (NAEP) (Mullis, 1980) as "a particularly good example of this approach...seminal in influencing our thinking" (Frederiksen and Collins, 1989, p. 29).

Frederiksen and Collins identify three crucial considerations that must be addressed in the design of a systemically valid testing system: the components, the standards to be sought in the design, and the methods by which the system seeks to encourage learning. The components of the system that Frederiksen and Collins seek to promote are:

> *Set of tasks.* The tests should consist of a representative set of tasks that cover the spectrum of knowledge, skills and strategies needed for the activity or domain being tested... The tasks should be authentic, ecologically valid tasks in that they are representative of the ways in which knowledge and skills are used in the real world.
>
> *Primary trait for each task and sub-process.* The knowledge and skills used in performing each task may consist of distinct sub-processes... Each subprocess must be characterized by a small number of primary traits or characteristics that cover the knowledge and skills necessary to do well in that aspect of the activity. The primary traits chosen should be ones that the test takers should strive to achieve, and thus should be traits that are learnable.
>
> *A library of exemplars.* In order to ensure reliability of scoring and learnability, it is important that for each task there be a library of exemplars of all levels of performance for each primary trait assessed in the test. The library should include exemplars representing the different ways to do well (or poorly) with respect to each trait... The library should be accessible to all, and particularly to the testees, so that they can learn to assess their own performance reliably and thus develop goals to strive for in their learning.
>
> *A training system for scoring tests.* There are three groups who must learn to score test performance reliably: (a) the administrators of the testing system who develop and maintain the assessment standards (i.e. the master assessors), (b) the coaches in the testing system whose role it is to help test takers to perform better, and (c) the test takers themselves, who must internalize the criteria by which their work is being judged. (p. 30)

In consideration of standards, Frederiksen and Collins identify:

> *Directness.* ...it is essential that whatever knowledge and skills we want test takers to develop be measured directly.
>
> *Scope.* The test should cover, as far as possible, all the knowledge, skills and strategies required to do well in the activity.
>
> *Reliability.* ...the most effective way to obtain reliable scoring that fosters learning is to use the primary trait scoring borrowed from the evaluation of writing.

> *Transparency*. The terms in which the test takers are judged must be clear to them if a test is to be successful in motivating and directing learning. In fact, we argue that the test must be transparent enough so that they can assess themselves and others with almost the same reliability as the actual test evaluators achieve. (p. 30)

Methods for fostering improvement on the test are:

> *Practice in self-assessment.* The test takers should have ample opportunity to practice taking the test and should have coaching to help them assess how well they have done and why...
>
> *Repeated testing*. If what is measured by the test is important to learn, then the test should not be taken once and forgotten. It should serve as a beacon to guide future learning.
>
> *Feedback on test performance.* Whenever a person takes a test, there should be a "rehash" with a master assessor or teacher. This rehash should emphasize what the testee did well and poorly on, and how performance might be improved.
>
> *Multiple levels of success.* There should be various landmarks of success in performance on the test, so that students can strive for higher levels of performance in repeated testing. (p. 31)

The paper *Complex, Performance-Based Assessment: Expectations and Validation Criteria* written by Linn, Baker, and Dunbar (1991) represents a watershed in the measurement literature as in it they propose a set of criteria for judgements that might be considered in the validation of performance assessments. What is notable about this paper is that the writers are respected members of the educational measurement community, and that the paper was published in a journal, *Educational Researcher*, which has a broad readership in North American faculties of education.

Linn, Baker, and Dunbar are concerned that arguments for the use of direct assessments, particularly those which focus upon the fidelity of the assessment tasks to the goals of instruction, place an unhealthy weighting on the face validity of the tasks. In stating that face validity alone is "not enough" they argue that

> evidence must support the interpretations and must demonstrate the technical adequacy of 'authentic' assessments...But what sort of evidence is needed, and by what criteria should these alternative to current standardized tests be judged? (Linn et al., 1991, p. 16).

Linn, Baker, and Dunbar find that few of the advocates of alternatives to standardized tests have addressed the issue of criteria for evaluating these measures. Linn, Baker, and Dunbar consider that a measure that is derived from actual performance or a simulation does not necessarily offer greater validity than a multiple-choice test. In response to their own statement of concern, Linn, Baker, and Dunbar propose a set of criteria that might be considered in evaluating the technical quality of performance assessments. They consider that these criteria might expand upon the base of "well established psychometric criteria for judging the technical adequacy of measures" (1991, p. 16). However, they caution that

> Reliability has been too often overemphasized at the expense of validity; validity has itself been viewed too narrowly. (Linn et al., 1991, p. 16)

In particular, Linn, Baker, and Dunbar review the historical criteria of efficiency, reliability and comparability of assessment from year to year, and speculate that these criteria would almost always favour the traditional multiple-choice assessment tasks in any comparison with the newer alternatives. They suggest that as there is expansion of modes of assessment there should also be an expansion of the criteria used to judge the adequacy of assessments. Linn, Baker, and Dunbar consider that modern views of validity, particularly as enunciated by Messick (1989b), present a theoretical rationale for such an expansion of the criteria. In broadening their view of validity to justify their criteria, Linn, Baker, and Dunbar caution that their "set of proposed criteria is not exhaustive" (1991, p. 16) and defend their position by claiming consistency with current views of validity and potential uses of new forms of assessment. The eight criteria proposed by Linn, Baker, and Dunbar for consideration in evaluating the adequacy of performance assessments are described here.

(1) Consequences – Linn, Baker, and Dunbar believe that the consequential basis of validity has come to the fore at an opportune moment for the advocates of authentic assessment[8]. They refer to Messick (1989b) and Cronbach (1988) as theoreticians who

> have stressed the criticality of giving attention to the consequential basis of validity prior to the recent pleas for authentic assessment

and find that

> consequences could rarely be listed among the major criteria by which the technical adequacy of an assessment was evaluated. (Linn et al., 1991, p. 17)

In particular, for performance-based assessments, Linn, Baker, and Dunbar argue that consequences must be given greater consideration:

> If performance-based assessments are going to have a chance of realizing the potential that the major proponents of the movement hope for, it will be essential that the consequential basis of validity be given much greater prominence among the criteria that are used in judging assessments. (Linn et al., 1991, p. 17)

Linn, Baker, and Dunbar point out that it cannot be assumed that all the consequences of performance-based modes of assessment will be positive or conducive to learning. They argue that there needs to be higher priority given to broadening the range of evidence collected about the assessment. In particular, the intended and unintended effects upon "the ways that teachers and students spend their time and think about the goals of education" need to be monitored (Linn et al., 1991, p. 17).

(2) Fairness – Linn, Baker and Dunbar believe that the criterion of fairness must be applied to any assessment, with specific judgements to depend upon the uses and interpretations of

---

8 The use of the term "authentic" in assessment stems from the work of Archbald and Newman (1988). These authors have refined the defining features of authentic achievement (Newmann and Archbald, 1992) to include (a) the production of knowledge, (b) disciplined inquiry, and (c) value beyond evaluation. Meyer (1992) argues that certain performance assessment tasks may not be authentic.

the assessment results. They identify the concern of biases against racial or ethnic minorities and cite results from the NAEP writing (assessed by open-ended essays) and reading (assessed mainly by multiple-choice questions) assessments where differences in achievement between Black and White students were similar for each type of test. Linn, Baker, and Dunbar state that these "gaps in performance among groups exist because of difference in familiarity, exposure, and motivation on the tasks of interest" (1991, p. 18). They advise that there needs to be substantial change in instructional strategies and resource allocation, particularly by providing training and support for teachers, to give students adequate preparation for assessments. In proposing these changes, they question the possibility of the success of such developments because:

> validly teaching for success on these assessments is a challenge in itself and pushes the boundaries of what we already know about teaching and learning. Because we have no ready technology to assist on the instructional side, performance gaps may persist. (Linn et al., 1991, p. 8)

These statements are puzzling in that they suggest that Linn, Baker, and Dunbar believe that methods of teaching directed at complex, time-consuming, open-ended assessments should be developed, with a goal of reducing the differences between the performance of Blacks and Whites; they appear to be suggesting that the problem is merely one of finding a "teaching technology". It is strange that Linn, Baker, and Dunbar do not raise the issue of gender-related differences in open-ended assessments. Perhaps this is because the authors have concentrated on assessments in reading and writing, where fewer significant gender differences have been observed compared with mathematics and science.

Linn, Baker, and Dunbar maintain that questions about fairness must also be asked of the procedures for scoring, particularly in the training of evaluators, to ensure consistent and unbiased ratings of performance. The authors examine the use of differential item functioning (DIF) procedures as a possible method of identifying items that may perform differently for minority groups. However DIF procedures require that several items be used as matching criteria in judging each individual item. This has not been feasible with

most performance assessments as the number of tasks has been quite small. As Linn, Baker, and Dunbar are writing from the perspective of consistency in measurement here, they appear to regret the inevitability of "greater reliance on judgemental reviews of performance tasks" (1991, p. 18).

Linn, Baker, and Dunbar express concern about prior knowledge as a source of bias. They cite reading assessments as being vulnerable, since children who have prior knowledge of the topic tend to show a higher level of comprehension. While this may be the case in assessment of general skills such as reading, it would be of concern in science only to those who espouse "process" assessments. A similar discussion concludes the section on fairness where Linn, Baker, and Dunbar cite Miller-Jones (1989) and the use of "functionally equivalent" tasks which would be "specific to the culture and instructional context of the individual being assessed" (Linn et al., 1991, p. 18). Miller-Jones recognizes that it would be "exceedingly difficult" to establish task equivalence. While Linn, Baker, and Dunbar consider that such development would pose a "major challenge" they do not give an opinion as to whether the challenge should be taken.

(3) Transfer and Generalizability – Generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Brennan, 1983) is identified by Linn, Baker, and Dunbar as providing "a natural framework for investigating the degree to which performance assessment results can be generalized" (1991, pp. 18-19). Generalizability theory is used to compare the magnitudes of variability for different factors of interest. Linn, Baker, and Dunbar consider that variability due to raters and sampling of tasks is the minimum to be considered. Results from direct writing assessments (e.g., Hieronymous and Hoover, 1987) and hands-on performance tasks in science (e.g., Baxter, Shavelson, Goldman, & Pine, 1990) show that the variance component for the sampling of tasks tends to be much greater than that for the sampling of raters. Linn, Baker, and Dunbar consider that this finding supports the research in learning and cognition that emphasizes the situation and context specific nature of thinking (Greeno, 1989). The authors consider that the limited

degree of generalizability across tasks must necessarily lead either to increasing the number of tasks or to the use of a matrix sampling approach.

Linn, Baker, and Dunbar argue that the traditional view of reliability is subsumed by this transfer and generalizability criterion. They identify a need for expansion of the traditional view of reliability:

> Consistency from one part of a test to another or from one form to another is insufficient. Whether conclusions about educational quality are based on scores on fixed response tests or ratings of performance on written essays, laboratory experiments, or portfolios of student work, the generalization from the specific assessment tasks to the broader domain of achievement needs to be justified. (1991, p. 19)

Linn, Baker, and Dunbar continue by calling for evidence that abilities demonstrated in specific modes of assessment are transferable to solving real-world problems. In their view, this should be a requirement of all assessments, multiple-choice as well as the newer alternative modes of assessment.

(4) Cognitive Complexity – According to Linn, Baker, and Dunbar this is a worthwhile criterion and should be used in judging all forms of assessment. They believe that many assumptions must be challenged, particularly in hands-on science and mathematics. For example, just because an assessment in science is hands-on, it does not necessarily incorporate problem-solving skills or more complex mental models. Similarly in mathematics they believe that complex open-ended problems do not always entail the use of complex cognitive processes. They declare that:

> Judgements regarding the cognitive complexity of an assessment need to start with an analysis of the task; they also need to take into account student familiarity with the problems and the ways in which students attempt to solve them. (Linn et al., 1991, p. 19)

It is interesting that Linn, Baker, and Dunbar recognize the importance of examining the student responses to questions in the analysis of cognitive complexity, rather than leaving the assessment of complexity solely to the inferences of "experts".

(5) Content Quality – This criterion specifies that the content of an assessment "be consistent with the best current understandings of the field" (Linn et al., 1991, p. 19). In addition, these authors consider it important that "tasks selected to measure a given content domain should themselves be worthy of the time and effort of students and raters" (p. 19). Linn, Baker, and Dunbar emphasize the value of involving subject matter "experts" in the design and review of tasks, particularly in the analysis of the quality of content knowledge displayed in various student responses.

(6) Content Coverage – Linn, Baker, and Dunbar appear to put lower priority upon content coverage. They describe it as "another potential criterion of interest" and state that:

> Performance assessment recognizes the importance of process sampling, giving it primacy over traditional content sampling. But breadth of coverage should not be overlooked. (1991, p. 20).

These comments do not appear to have been directed towards any specific area of assessment. While these assertions might be accepted without too much debate in assessments of written comprehension or reading, in science there has been much dispute over the focus of performance assessment. Linn, Baker, and Dunbar observe that one of the consequences of gaps in content coverage is the under-emphasis of parts of the content that are excluded from the assessment. They believe that this is an area where there may have to be some trade-off with other criteria. In fact, they suggest that this may be a place where "traditional tests appear to have an advantage over more elaborate performance assessments" (Linn et al., 1991, p. 20).

(7) Meaningfulness – This criterion addresses the issue of provision of more worthwhile educational experiences. Linn, Baker, and Dunbar consider that some "investigation of student and teacher understandings of performance assessments, and their reactions to them would provide more systematic information relevant to this criterion" (1991, p. 20). They speculate that low performance on NAEP assessments might be related to students' perceptions of lack of meaning in the test situation.

(8) Cost and Efficiency – Linn, Baker, and Dunbar compare labor-intensive performance assessments to paper-and-pencil, multiple-choice tests. They argue that greater attention must be given to the "development of efficient data collection and scoring procedures" for performance assessments (Linn et al., 1991, p. 20).

Linn, Baker, and Dunbar conclude with a reminder that their eight criteria are not an exhaustive set. Their key point is that "the traditional criteria need to be expanded to make the practice of validation more adequately reflect theoretical concepts of validity" (Linn et al., 1991, p. 20). In addition, they assert that this perspective is not a "theoretical nicety" but a way of identifying suitable criteria to make judgements about the relative merits of newly developed alternative assessments. To those who might accuse them of "stacking the deck in favor of alternative assessments over traditional fixed-response tests" Linn, Baker, and Dunbar respond that:

> the issue is not which form of assessment may be favored by a particular criterion, however. Rather, it is the appropriateness and importance of the criteria for the purposes to which the assessments are put and interpretations that are made of the results. (1991, p. 20)

Linn, Baker, and Dunbar consider that it is essential that the evolving conceptions of validity should be dominant in any argument about traditional and alternative forms of assessment, particularly in the context of the "fundamental purpose of measurement – the improvement of instruction and learning" (1991, p. 20).

Another group of authors, Lane, Parke, and Moskal in *The Principles for Developing Performance Assessments* (1992), refer to Messick's work as fundamental in their conceptualization of validity. They cite Frederickson and Collins (1989), Glaser (in press), Linn, Baker, and Dunbar (1991) and Merhens (1992) as references for establishing criteria to ensure reliable and valid performance assessments. The *Principles* were developed in the context of the QUASAR (Quantitative Understanding: Amplifying Student Achievement and Reasoning) project, an American instructional program based at the University of Pittsburgh, that attempts to promote the acquisition of thinking and reasoning

skills in mathematics. Lane and her co-authors suggest that their principles "may be extended to other types of performance assessments and can be used with other subject matters" (1991, p. 2). Twenty-six principles are grouped within the following categories:

construct specification

content representation and relevancy

specification of task format and task representation

task wording and directions

task fairness

specifications for the scoring rubrics

These categories, and the attendant principles, are intended for sequential use in the planning and implementation of performance assessments. While Lane, Parke, and Moskal claim that these principles are "discussed in relation to current conceptualizations of validity" (1992, p. 3) there is a paucity of explicit discussion in this paper about the consequential aspects of validity.

The chapter "Evaluating Test Validity" by Shepard in *Review of Research in Education* (1993) also builds upon the work of Messick (1989a; 1989b). Shepard covers the historical aspects of validity in the first section, and uses the second to reconfirm "construct validity as the whole of validity" (Shepard, 1993, p. 405). The third section is devoted to an analysis of Messick's unified theory, followed by Shepard's reformulation of Messick's theory in terms of evaluation arguments. Shepard presents specific examples and concludes the chapter with a discussion of the implications for the development of the next set of standards for educational and psychological testing[9]. The pertinent sections for this review are Shepard's reformulation of validation in terms of an evaluation argument, and her discussion of the potential for revision of the standards.

---

9 The planning for these standards has started with the formation of a joint committee of the NCME, AERA and APA. The draft standards are to be reviewed during 1995.

Shepard considers that Messick's presentation of validity in the four-fold table[10] – shown as Figure 2 below – has the potential for a "new segmentation of validity requirements" (1993, p. 426).

	TEST INTERPRETATION	TEST USE
EVIDENTIAL BASIS	Construct validity	Construct validity + Relevance/utility
CONSEQUENTIAL BASIS	Value implications	Social consequences

Figure 2. Messick's Facets of Validity Framework (Messick, 1989b)

Shepard expresses three concerns about this schemata:

> (1) The faceted presentation allows the impression that values are distinct from a scientific evaluation of test score meaning.
> (2) By locating construct validity in the first cell and then reinvoking it in subsequent cells, it is not clear whether the term names the whole or the part. I argue for the larger, more encompassing view of construct validity.
> (3) The complexity of Messick's analysis does not identify which validity questions are essential to support test use. (Shepard, 1993, pp. 426-7)

Shepard believes that the use of positivistic principles, where facts and values are separated, is "no longer defensible in contemporary philosophy of science" (1993, p. 427). While Shepard recognizes Messick's acknowledgment that "scientific observations are theory-laden and theories are value-laden" (Messick, 1989b, p. 62), she argues that Messick lapses on his next page "into discourse that separates value implications from substantive or trait implications". Shepard states that this:

> plays into the hands of researchers who deny that their construct definition or predictive equation follows from value choices. This should not be read to mean that scientific facts are indistinguishable from value judgements. Although scientific inquiry is distinct from politics and moral philosophy at the extremes, the concern here is with value perspectives that are entwined with scientific investigations. (1993, p. 427)

---

10 This is the table used in the *Handbook of Educational Measurement*. I chose to use the similar table from the *Educational Researcher* article earlier in this study.

Shepard also voices concern with Messick's sequential arrangement of cells which leads to an apparent segmentation of validity. While Messick describes his framework as a "progressive matrix" with construct validity appearing in every cell, and something more to be added in each subsequent cell, Shepard identifies a significant problem with this approach:

> Messick has implicitly equated construct validity with a narrow definition of score meaning, whereas I would equate it with the full set of demands implied by all four cells, which all involve score meaning. Intended effects entertained in the last cell are integrally a part of test meaning in applied contexts. (1993, pp. 427-8)

Shepard clarifies that her disagreement with Messick relates more to issues in communication than to divergent perspectives about validity. Shepard emphasizes that it is vitally important to communicate issues that have arisen to those involved in the theory and practice of measurement. She refers to how Cole and Moss (1989) had earlier used the term *validity* to refer only to the interpretive component of a framework but:

> since then, we have expanded our definition of validity to include the consequential component, in part, because we were concerned that excluding consideration of consequences from the definition of validity risks diminishing its importance. (Moss, 1992, p. 235)

Shepard goes on to look at the operational level of planning and conducting validity evaluations. She argues that Messick's sequential approach "may misdirect the conceptualization of theoretical frameworks intended to guide validity evaluations" (Shepard, 1993, p. 429). Shepard contends that "measurement specialists need a more straightforward means to prioritize validity questions" and argues that the current *Standards* (AERA, APA, & NCME, 1985) provide little help. The absence of a coherent conceptual framework in which to organize validation does not help the standards to:

> answer the question "How much evidence is enough?" nor do they clarify that the stringency of evidential demands should vary as a function of potential consequences. (Shepard, 1993, p. 429)

In response to this problem, Shepard proposes that:

> validity evaluations be organized in response to the question "What does the testing practice claim to do?" Additional questions are implied: What are the arguments for and against the intended use of the test? and What does the test do in the system other than what it claims, for good or bad? All of Messick's issues should be sorted through at once, with consequences as equal contenders alongside domain representativeness as candidates for what *must* be assessed in order to defend test use. (1993, pp. 429-30)

Shepard credits Cronbach (1988, 1989) for proposing that validation should be considered as an evaluation argument. In this process, the evaluator identifies relevant questions for intensive research, justifying this particular set in terms of the "prior uncertainty, information yield, cost and leverage"[11] (Shepard, 1993, p. 430). The process is taken further by Kane (1992) who conceptualizes validation as an interpretive argument. Shepard paraphrases Kane's criteria for evaluating the argument as:

> (a) The argument must be clearly stated so that we know what is being claimed; (b) the argument must be coherent in the sense that conclusions follow reasonably from assumptions; and (c) assumptions should be plausible or supported by evidence, which includes investigating plausible counterarguments. (Shepard, 1993, pp. 430 - 31)

Thus by setting out assumptions it is possible to identify specific areas of study and the types of evidence that are needed to support specific hypotheses.

Shepard concludes her chapter by confirming that "construct validation is the one unifying and overarching framework for conceptualizing validity evaluations" (1993, p. 443). Both analysis of test content (content validity or representation) and empirical confirmation of hypothesized relationships (criterion validity) are essential but not sufficient for Shepard in her approach to construct validation. She insists that there must be a conceptual framework which:

---

11 Shepard identifies "leverage" as referring to the importance of the study information in achieving consensus about test use in the relevant audience.

> portrays the theoretical relationships believed to connect the test responses to a domain of performance and desired ends implied by the intended test use. In all but rarefied research contexts, test uses have intended consequences that are an essential part of the validity framework. Given that theory testing must also include empirical evaluation of the most compelling rival hypotheses, construct validation entails a search for both alternative meaning and unintended consequences as well. (Shepard, 1993, p. 443)

Shepard encapsulates these developments by considering a shift in the key validity question from "Does the test measure what it purports to measure?" to "Does the test do what it claims to do?" (Shepard, 1993, p. 444). The metaphors that Shepard uses illustrate the change in emphasis – from "truth in labelling" to searching for both side effects and intended effects of newly developed drugs. Too many "side-effects" would necessarily lead to a decision to remove the test from that specific application. Shepard's final paragraph contains a reminder there is no universality in time or location for test use; therefore all validation studies must be considered context-specific. Shepard suggests that test developers should:

> be able to specify the evaluation argument for the test use they intend and gather the necessary evidence, paying close attention to the competing interpretations and potential side effects. (1993, p. 445)

In developing an evaluation argument for performance assessment in science, the choice of construct labels is an important factor, and will be the focus of the next part of this literature review.

PROCESSES OF SCIENCE

My focus in this study is the validation of hands-on performance assessments in the context of school science. In arguing that performance assessment in science poses special challenges, specific concerns arise from the conceptual framework implied by the use of "processes of science" as an organizing theme. The construct labels that are used as a result of this framework have had a significant impact on the design of curriculum and assessment in the English-speaking world. The "processes of science" issue must play a

significant role in any discussion of hands-on assessment of student performance in science. Donnelly and Gott examine the impact of this approach on curriculum theory and development, and on large scale curriculum evaluation and student assessment. They conclude that:

> process specification, the cognitive and epistemological status which has been ascribed to them, and their place in the formal structures which have been used to describe the science curriculum, have generally remained problematic. (1985, p. 237)

In an attempt to resolve some of these problems, Millar and Driver (1987) consider many of the pertinent issues in a paper entitled *Beyond Processes*. They start by tracing the recent use of the term 'process' back to the work of Gagné, and his influence upon the course *Science - A Process Approach* (1965). Gagné argues that scientific concepts and principles are acquired through the operation of basic scientific processes (such as observing, classifying and inferring) and integrated processes (such as formulating hypotheses, interpreting data, etc.). Gagné believes that processes are skills used by all scientists, and can be learned by students; they are transferable across content domains. Donnelly and Gott are concerned about the lack of attention "devoted to their [processes] definition, interrelationship or justification" (1985, p. 237). Millar and Driver address these concerns in *Beyond Processes* (1987). Millar and Driver discuss the apparent separation of content and process in science education. A teaching approach that stresses "content" is seen as transmissive and passive, whereas stress on "process" is seen as more active and progressive. They support Black who argues that:

> the content-process separation is artificial. One does not observe except through selective attention, guided by one's aims and one's theories. One cannot tackle even the simplest experimental problem without some model of the working of the system under investigation. (Black, 1986, p. 672)

Millar and Driver present a multi-dimensional critique of the claims for process science. They start with an examination from the perspective of philosophy of science where the processes of science are seen as the methods of science, and of scientists. Feyerabend

(1975), perhaps the most uncompromising of modern philosophers writing about methods, is cited as providing the argument for "the underlying point that there is no algorithm for gaining or validating scientific evidence" (Millar and Driver, 1987, p. 40). Donnelly and Gott (1985) suggest that conceptualizing science in terms of the activities of professional scientists is a top-down approach, providing processes that are "vague and unoperationalized" (p. 239).

Millar and Driver continue their critique from the perspective of cognitive psychology. The view that learners are actively constructing personal knowledge has a relatively long history; Millar and Driver (1987) identify the work of Piaget, Bruner and Kelly as examples of this approach. Millar and Driver indicate that contemporary perspectives on cognition (e.g., Bereiter, 1985; Resnick, 1983)

> reject the view that learning is a one-way process whereby the learner receives and organizes stimuli from an external world. Instead learning is viewed as an active process in which the learner brings prior sets of ideas, schemes or internal mental representations to any interaction with the environment. (Millar and Driver, 1987, p. 45)

This view gains support from the "alternative conceptions" literature (Driver, Guesne, and Tiberghien, 1985; Osborne and Freyberg, 1985): the mental representations that learners bring to a situation enable them to make predictions and conceptualize physical phenomena in ways that are internally consistent, but are not accepted as "science".

In examining the impact of a process approach on pedagogy, Millar and Driver are concerned about the implicit expectation of transferability:

> The idea of teaching the processes of science implies both that general cognitive skills can be transferred from the specific area and context within which they are taught to other analogous contexts, and that instruction can foster the learner's progress or development in deploying these skills. (Millar and Driver, 1987, p. 51)

Millar and Driver doubt both that such content-independent processes can be taught and that transfer will occur to new situations. They posit that learners tend to understand new situations through reasoning by analogy with other previously encountered situations,

rather than by any general procedures. Such use of analogy or modeling, almost through trial and error, enables learners to transfer and generalize from one context to another. Millar and Driver also question the possibility of a "progressive scale of difficulty or of the pattern of stages of children's performance" (1987, p. 53) and express doubt that such a scale could be adequately described. In comparison, analysis of logical structures of subject matter has led to "an empirical base of knowledge about the development of patterns of children's growth in conceptual understanding" (Millar and Driver, 1987, p. 53). Donnelly and Gott (1985, p. 239) also examine the "strength of arguments for generalized processes when confronted with the authority of the discipline-based curriculum" and find them to be weak; they believe that this weakness should be acknowledged in order that science processes coexist with the traditional science discipline in secondary curricula[12].

Processes in Hands-on Science Assessment

Donnelly and Gott (1985) offer two main explanations why science processes have become such a major feature of large scale assessments in science:

> the objective of assessment in science, which compels a search for a unifying aspect of science disciplines; and
>
> the need for assessment to extend across pupils and institutions, and thus to minimize the impact of curriculum-based variables. This assumes that curricula are largely based on traditional content. (p. 240)

These reasons ring particularly true as the struggle to present a unified face for science from primary to secondary grades continues in England with the National Curriculum

---

12 Donnelly and Gott offer their working definition of 'science processes':

classes of tasks undertaken by pupils which

- can be identified across a wide range of disciplinary areas; and
- can be systematically connected with a specifically scientific epistemology, which is analytical, manipulative and materialist. Its proximate results are variable-based descriptions of phenomena (data) and the establishment of functional relationships between variables. (1985, p. 239)

(Department of Education and Science and the Welsh Office, 1991) and in the U.S.A. with the proposed National Science Education Standards (National Committee on Science Education Standards and Assessment {NCSECA}, 1993). Both of these national projects have an embedded emphasis upon the processes (or skills) of science. In England this is identified as Science Attainment Target 1 - Scientific Investigation, while in the U.S.A. this is seen as a set of inquiry skills that students "should be able to demonstrate in a new experiment" (NCSECA, 1993; p. 56). As there are also content dimensions for each of these national projects, neither can be considered set exclusively in a "science processes" mode. However, the Techniques for the Assessment of Practical Skills in the Foundation Sciences (TAPS) Project (Bryce, McCall, MacGregor, Robertson, & Weston, 1984) was so conceived. Bryce and Robertson (1985) describe the focus of TAPS as "the non-trivial aspects of practical skills as they manifest themselves in the classroom" and state that the skills are "conceptualized in accord with the ways in which teachers think and act in the laboratory". It is of interest that these authors focus on the teacher, rather than on the students who will perform the tasks, reflecting their interest in enabling teachers to feel that the assessment scheme supports their work. The framework used in the TAPS scheme is based upon the work of Whittaker (1974) and Swain (1974). The TAPS team has been extremely thorough in the definition of skills, and has developed what is called "a step-up approach". Basic skills (e.g., observation, recording and measurement) are assessed with tasks from a bank of over 300 items. The higher level skills (e.g., inference and selection of procedures) constitute the second level, with students performing investigations as the pinnacle of achievement. The TAPS team is explicit that students should start at the basic level, and work through a hierarchy of process tasks until they are capable of attempting investigations (Bryce et al., 1984).

This "step-up" theory of "science processes" in assessment has been strongly criticized by Woolnough and Allsop (1985), Woolnough (1989, 1991), Millar and Driver

(1987) and Millar (1989, 1991). Woolnough argues against a reductionist approach with:

> a tightly prescriptive structure for science teaching which has reduced itself to a series of small component parts. Though not claiming each of these parts, in itself is scientifically significant the implied message is that when they are ultimately put together by the pupil they will produce competence in the 'scientific method'. (1989, p. 41)

Woolnough is concerned that using discrete processes in assessment will lead to separate teaching of different skills. Woolnough values complete investigations in science and argues for a holistic approach where students "learn to do investigations by actually doing scientific investigations, simple ones at first but complete investigations none the less, becoming more sophisticated as confidence and experience increase" (Woolnough, 1989, p. 43). Woolnough wrote his chapter before the publication of the National Curriculum for Science in December 1988 (Department of Education and Science and the Welsh Office). In a footnote, Woolnough rejoices at the holistic approach that is presented for "Attainment Target 1: Exploration of Science" in the 1988 curriculum. But in 1991, a revised curriculum was published, and "Attainment Target 1: Scientific Investigation" appears to have been re-aligned with a step-up approach.

Millar and Driver (1987) are particularly critical of the process approach to assessment. They identify two reasons for concern: (1) the influence of content and context upon student scores, and (2) the problem of identifying processes used from the written outcome, as tasks do not necessarily predetermine strategies that students use. This latter point is significant, particularly with respect to complex tasks. The requirement that tasks be uni-dimensional is an important aspect of the TAPS program (Kempa, 1986). Bryce and Robertson (1985, p. 18) describe as "purifying" the modification of tasks to ensure that they focus upon only one specific objective. The TAPS program in Scotland was developed in parallel with the Assessment of Performance Unit (APU) for science in England and Wales. The work of the APU has been the more influential of these two British-based assessment projects and deserves detailed examination.

The Assessment of Performance Unit for science (Johnson, 1989) examined performance in science of English and Welsh students age 11 (Russell, Black, Harlen, Johnson, & Palacio, 1988), age 13 (Schofield, Black, Bell, Johnson, Murphy, Qualter, & Russell, 1988) and age 15 (Archenhold, Bell, Donnelly, Johnson, & Welford, 1988). The APU work provided a significant foundation for the first attempt at practical-based assessment in North America – the National Assessment of Educational Progress (NAEP) project *A Pilot Study of Higher Order Thinking Skills Assessment Techniques in Science and Mathematics* (Blumberg, Epstein, MacDonald, & Mullis, 1986). APU-derived tasks appeared in the New York State Grade 4 performance tasks in 1989 (New York State Department of Education), which itself was influential in the design of the fifth-grade tasks used by the California State Department of Education in 1990.

Johnson (1989) reports that the APU inherited a process-based framework from a Science Working Group, although it was acknowledged that "processes are rarely deployed without the simultaneous use of knowledge and conceptual understanding" (Johnson, 1989, p. 7). The first task of the APU Science Monitoring Team[13] was to develop an appropriate approach to the issue of process-content interdependence. As England had no national curriculum at that time, content lists were developed and refined by consultation with teachers. The content was grouped into the three disciplines of biology, chemistry and physics. Process skills are represented by a framework of Science Activity Categories, each considered to portray an identifiably different activity.

---

13 The APU teams were based at two centres, at Kings College, London (formerly Chelsea College) for ages 11 and 13, and at the University of Leeds for age 15.

The categories are:

> Using symbolic representations
> Use of apparatus and measuring instruments
> Using observations
> Interpretation and application
> Design of investigations
> Performing investigations

Johnson (1989, p. 10) reminds us that "it has never been claimed that these Categories are mutually exclusive" and that there is clearly some overlap in terms of skills and abilities. The Monitoring Teams chose to use a practical mode of assessment for "use of apparatus and measuring instruments", "using observations", and "performing investigations", with pencil and paper tests for the other categories. Circuses of experiments were chosen for both "use of apparatus and measuring instruments" and "using observations"[14]. "Performing investigations" was structured so that a student would spend an extended period of time working through a single investigation. The tasks and the procedures developed by the APU make up perhaps the most rich legacy for those following in the design and implementation of performance assessment in science. Welford, Harlen and Schofield describe how tasks were chosen for the "using observations" circus:

> In **observation**, between 40-50 questions are selected at random from the bank for each survey at each age. These are divided into two or three circuses of experiments, each administered separately. Between 12 and 20 experiments are distributed among eight or nine stations in the laboratory or classroom being used. Time is the basis for distributing questions among the stations. Extensive question trailing has resulted in tasks which require two, four, six, or eight minutes, and so questions are grouped to allow completion of a station within eight minutes. (1985, p. 25)

---

14 The circus approach involves students working through a set of as many as 8 or 9 different experiments, called stations, one after another. The name circus was given because the students move around, generally in some kind of circular pattern.

This large number of questions, and a similar set for "use of apparatus and measuring instruments", enabled the APU to collect annual data from 1980 through 1984. The APU produced 15 extensive reports, and 11 short reports for teachers.

While the circus approach to student practical work was a regular feature in the practice of science education in England, open-ended investigations most certainly were not. Gott and Murphy, in *Assessing Investigations at Ages 13 and 15* (1987), wrote of the position of processes in school science:

> In general, little emphasis is placed on what have become labelled the 'processes' of scientific enquiry; planning, interpreting and so on. These 'processes' are seen as serving the ends of concept acquisition rather than being of intrinsic value in themselves. This is not to argue that such 'processes' are absent from this view of science: rather it is to suggest that they are often unacknowledged and, where present, are there because of the style of teaching rather than explicit aims of the course.
>
> Assessment, not unnaturally, has similar aims. It is concerned with pupils' ability to explain phenomena using their conceptual knowledge and understanding. (p. 6)

Gott and Murphy go on to describe courses designed to develop the processes of science. They warn that:

> The danger in such approaches lies in the degree to which the emphasis on concept in the one case or 'process' in the other effectively underplays or even denies the significance of the other. (p. 6)

Investigations were seen by the APU monitoring teams as catering deliberately to the interaction between process and concept by presenting the students with practically based problems. As much of the language used by teachers and those concerned with assessment has both everyday and technical meaning, the APU chose to be explicit and defined a **problem** as:

> a task for which the pupil cannot immediately see an answer or recall a routine method for finding it. (Gott and Murphy, 1987, p. 7)

The set of problems that was developed by the APU was by necessity different from the traditional exercises that students meet in school science. These investigations were

designed so that "science concepts and procedures are essential elements of any solution" (Gott and Murphy, 1987, p. 8). Gott and Murphy report that a descriptive framework for the categorization of problems emerged as the trials proceeded. They point out that these descriptions are based upon:

> what an assessor would consider to be important and does not reflect either the pupils' perception of the problems or the knowledge they, individually use to solve them. (Gott and Murphy, 1987, pp. 8-9)

The elements of the APU investigation framework are **purpose**, **nature**, **content** and **context**. The **purpose of a problem** is seen as a statement of what the student is asked to do. Three types of purpose are identified: "*Decide which...*" problems which may lead to students designing investigations to make a choice between competing products; "*Find a way to...*" problems which ask students to find a way to extend the use of apparatus beyond its usual; "*Find the effect of...*" problems which are usually set in a more scientific context and are concerned with the effect of changing one or more variable. The **nature of a task** is identified as the arrangement of apparatus, and the development of a single strategy or a series of interconnected investigations. The **content** and **context** are similar in that either may be scientific (i.e., likely to have been encountered only in the school laboratory) or may be everyday (e.g., concerned with the absorbing properties of paper towels).

These elements suggest limits upon the **conceptual understanding** and the **procedural understanding** that might be expected or applied in any investigation. **Conceptual understanding** is seen as a "loose division of concepts into taught science and everyday science, or some combination of these" (Gott and Murphy, 1987, p. 12). In addition, the APU perceives that "conceptual understanding is vital in the identification of the key variables in a problem", particularly in controlling variables for a fair test (Gott and Murphy, 1987, p. 12). **Procedural understanding** is identified as "strategies of scientific enquiry such as will occur in a many variable problem" (Gott and Murphy, 1987,

p. 12). The specific actions that constitute a scientific procedure, according to the APU, are shown in Figure 3.

- defining the status of variables as:
  - independent
  - dependent
  - or control
- systematically varying the independent variable,
- developing a measurement strategy for the dependent variable,
- controlling variables by choice of apparatus,
- choosing an appropriate scale for the quantities of variables, This will require a consideration of the match between quantities and the measurement instruments available,
- developing a strategy for sifting complex data involving many variables,
- transforming data from one form to another.

During the course of the investigation a variety of intellectual and practical skills may also be called upon:

- setting up apparatus
- reading instruments
- recording data
- interpreting data
- evaluating the data obtained or the methods used.

Figure 3. APU Procedures for Scientific Enquiry
(Gott and Murphy, 1987, p. 15)

The APU used this framework to generate and explore problems (Gott and Murphy, 1987). The approach developed by the APU for assessment of student performance in investigations involves a trained observer completing a checklist developed previously by the monitoring team. The checklist enables the observing teachers and the monitoring team to reconstruct the investigation in the manner the student performed it. Data collected from the checklists enable the APU to provide aggregate results to describe student performance. The APU developed holistic scales of performance with:

> criteria based on a scientist's view of what is or is not an appropriate experiment. We have not attempted to say, for instance, that a particular experiment is good for that age and ability of pupil but rather to say – this is what a scientist would do – how does a pupil match up to this 'expert'. (Gott and Murphy, 1987, p. 33)

The researchers report that:

> The data collected so far suggest that many pupils can make a very creditable attempt at investigations which emphasize the 'processes' and 'procedures of science'. (Gott and Murphy, 1987, p. 40)

Compared with traditional approaches to assessment using written questions, where the results uniformly indicate low levels of performance, the "higher success rate of the investigative problem must come as something of a surprise" (Gott and Murphy, 1987, p. 40). Investigations such as these were not commonplace in secondary schools, yet students appeared to approach the tasks with confidence and success. With regard to age-related differences in performance, Gott and Murphy (1987) report the somewhat unexpected finding that pupils:

> do not seem to improve significantly between the ages of 13 and 15, a finding which suggests either that pupils are already performing to their limit at age 13 or that the science of years 4 and 5 is doing little to assist their problem-solving capacity, at least on this type of problem[15]. (p. 40)

Gott and Murphy suggest five possibilities which may explain these findings:

1. ***The reduced writing load***: it was found that the sub-sample of pupils who were asked to 'write-up' their experiment did so in a fashion which conveyed much of what they had done. Students were capable of writing the report, if asked to do so.

2. ***The practical nature of the task presents many more clues which assist in the perception of the problem; it is the failure to grasp the problem in the first place which is the difficulty***. Gott and Murphy describe the differences in performance on prose, pictorial and practical versions of the paper towel

---

15 The English system of the time identified years 4 and 5 as students in their 4th and 5th years of secondary schools, equivalent to North American grades 9 and 10.

investigation[16]. In the prose version, pupils are presented with a written stem and no description of the possible apparatus; the pictorial version presents the pupils with a picture of possible apparatus and an identical stem, and the practical version has a complete set of equipment for the pupils to use in working through the experiment, along with the written question stem. In the APU study, the percentage of students performing at the highest level was lowest for prose (12%), higher for pictorial (24%) and highest (43%) for practical versions. At the lowest level of performance the percentage for prose was 71%, pictorial 51%, and 23% for the practical version. This pattern of data was interpreted as indicating that "practical tasks allow more pupils to see the problem in its entirety rather than as a series of disconnected activities". Gott and Murphy make the suggestion that "interaction with the apparatus **during** the experiment is an important facet of practical work" (Emphasis in original. p. 45).

3. ***The practical nature of the task provides opportunity for reconsideration and refinement of the experiment, an interaction which permits pupils to identify the necessary stages***. Gott and Murphy describe how the APU broke up investigations into component parts for prose, pictorial and practical versions. The performance of students was generally lower in the versions which required extended responses. Gott and Murphy argue that:

> practical tasks (and to a lesser extent extended response questions) not only allow pupils to perceive the questions as a whole, but also give some measure of feedback as they attempt to put together a logical sequence of activities that will lead to a solution. This feedback is either very abstract and self generated or, more likely, non-existent in the structured questions. (1987, p. 49)

---

16 In this investigation pupils are given three different kinds of paper towel and asked to "Describe an experiment you could do to find out which kind of paper will hold the most water" (Gott and Murphy, 1987, p. 43).

4. ***The tasks require very little in the way of taught science concepts; their absence is responsible for the improved performance:*** Gott and Murphy point to detailed analysis of student performance over several different investigations where conceptual understanding, either in the form of conceptualization of the problem or appreciation of the effects of specific variables, is seen as a limiting or enabling factor. Gott and Murphy conclude that:

> the activities that we have labelled problem-solving in his booklet are characterized by the **mutual dependency** of scientific procedural and conceptual understanding. From this viewpoint the pupils are regarded as having to access the pool of concept and knowledge available to them in order to **first conceptualize** the problem. A procedural strategy has then to be developed but the pupils' particular conceptualization of the problem is the link which **determines** the procedures that are understood to be appropriate in the problem situation. In practical activity the pupils can both refine their procedural strategy and their conceptualization of the problem by evaluating their own procedures and their outcomes in situ. (Emphasis in original. 1987, p. 51)

5. ***The everyday context in which the tasks are set encourages more pupils to "have a go".*** The APU considered this to be an important issue and ran a special study in which student performance was scrutinized across a range of investigations (10 for most students). Examination of these results indicates that "questions set in a scientific context can inhibit some pupils' performance" (Gott and Murphy, 1987, p. 51). The authors argue that the effect appears to be linked with "pupils' belief in their own incompetence in a specific topic area". Inconsistency in pupil performance may be explained in part by pupils perceiving:

> alternative problems which appear to be loaded with the concepts they associate with the content: they can therefore decide that they do not 'know' how to solve such a problem. (Gott and Murphy, 1987, p. 51)

One effect of setting the problem in an everyday context was to lead "some pupils to decide an 'everyday' answer is all that is required" (Gott and Murphy, 1987, p. 51). A corollary to this is a cause for concern:

> The idea that we should only behave scientifically when the task looks scientific should give us cause to reflect on the nature of science teaching. If we only ever carry out investigations related to the concepts of science using apparatus that is never encountered outside the laboratory, there is perhaps little wonder that pupils assume that scientific methods can only be applied in such circumstances. (Gott and Murphy, 1987, p. 51)

This issue of context was identified as a focus for a subsequent phase of APU Science research – unfortunately cancelled because government funding for the project was terminated.

By 1987, the educational climate in England and Wales had become relatively unsettled. The four years which led to the 1988 introduction of the General Certificate of Education (GCSE) were followed by lobbying leading to publication of the National Curriculum for Science (Department of Education and Science and the Welsh Office, 1988). In their analysis of the implications of the APU results, Gott and Murphy refer to the significance of using the view of science adopted by the APU to guide an assessment program, particularly as "the ability to plan and perform investigations is an important part of science education for all pupils" (Gott and Murphy, 1987, p. 52). Gott and Murphy consider that accepting such a value leads to a need for clarification of purpose, and types of investigation, together with a formalization of the role of procedural understanding in science education. They perceive that:

> the interaction of procedure and concept inherent in such problem-solving activities is a useful addition to current courses and will allow pupils to demonstrate their conceptual understanding in a situation which does not rely solely on explanation. (Gott and Murphy, 1987, p. 53)

My examination of the APU's work has concentrated upon the role of conceptual and procedural knowledge in investigations. The **Performance of Investigations** was identified by the APU as the "putting together of the component activities within science performance, hence an overlap with other categories was assumed" (Murphy, 1989, p. 149). For the other categories, assessed by stations or written questions, there was a self-imposed requirement to identify a unique category for each question:

> in establishing category definitions questions were written to *fit* one category only. The degree of *fit* was established during the external validation exercise carried out by groups of experts in the education field. The criterion of fit used was the *burden* of the given question's demand, i.e. its ultimate loading in terms of the demands made upon the pupils. Questions which failed to fit the category they were designed for were rejected or rewritten. If this criterion was met then what might be overlap in principle would in practice have little effect on performance score correlation. (Italics in original. Murphy, 1989, p. 149)

This was not achieved. Murphy reports that the APU came to identify **Use of graphical and symbolic representation** (Category 1) and **Use of apparatus and measuring instruments** (Category 2) as "enabling skills" which contribute to student performance in other categories. Murphy also relates that the data gathered over the five years of surveys indicate that:

> many questions made demands on pupils other than those specified in the definition of the question type. The question type was seen to represent the major demand of a question but the subsidiary demands also appeared to influence pupil's performance. (1989, pp. 149-50)

Johnson (1989) describes some of the pilot work that was done to provide empirical, albeit correlational, evidence of relationships among categories. She warns that

> empirical data cannot substitute for educational judgment. A poor correlation between performance on two groups of questions might support an assumption that the two groups measure different things, or might throw doubt on a previous assumption about equivalence. A strong association between different test questions is not sufficient evidence that these do measure the same thing(s). (Johnson, 1989, p. 13)

Correlations between subcategory test scores were found to be in the range of 0.55 to 0.75, with modal values in the order of 0.6. These correlations appear to have been calculated between the aggregated scores for each test package. Johnson is particularly tentative in discussing these "moderately high correlations" which "could be taken as confirmation of the overlap between subcategories" or "might also be merely coincidental" (1989, p. 13).

The APU also pioneered some significant technical procedures in the establishment of reliability and generalizability of student performance. An explanation of some of these

procedures will follow in the section examining reliability and generalizability of student performance.

RELIABILITY – MEANINGS AND REQUIREMENTS

The widely accepted guidelines for the use and development of tests, the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1985), cover technical standards for test construction and evaluation as well as professional standards for test use. The section on reliability is written in the language of classical reliability theory — standard error of measurement, reliability coefficient and true score. Classical reliability theory has definite limitations as a model when applied to performance assessments. These problems are identified below and the use of Generalizability Theory (Shavelson and Webb, 1991) is explored.

Classical Theory

Feldt and Brennan (1989) describe the historical perspectives of reliability and classical reliability theory as an attempt to quantify the consistency and inconsistency in examinee performance. They define the standard error of measurement as:

> the standard deviation of a hypothetical set of repeated measurements on a single individual. Because such measurements are presumed to vary only because of measurement error, the standard deviation represents the potency of random error sources. (p. 105)

Feldt and Brennan characterize the reliability coefficient as quantifying "reliability by summarizing the consistency or inconsistency among several error-prone measurements" (1989, p. 105). Error manifests itself by depressing the correlation coefficient between two experimentally independent measures. These two statistics are seen by Feldt and Brennan as having specific limitations and advantages in use. Standard error is useful as the error is presented in score-oriented units; it can be used to suggest how testing procedures can be improved, and it is relatively stable from group to group. However the

standard error cannot be compared from one instrument to another, or used to compare different scoring procedures for the same instrument. Reliability coefficients have a much greater generalizable use. For example, they can be compared between instruments — instruments with coefficients of less than 0.70 are considered unsuited for individual student evaluations (Feldt and Brennan, 1989). Feldt and Brennan warn that reliability coefficients are sensitive to the character of the group from which the data were collected. Feldt and Brennan consider that "true score" is a further vital concept in quantification of reliability. Their definition is taken from conventional practice:

> the true score of a person is regarded as a personal parameter that remains constant over the time required to take at least several measurements, though in some settings a person might be deemed to have a true score subject change almost moment to moment. (1989, p. 106)

In warning against considering the true score as merely the "limit approached by the average of observed scores as the number of these observed scores increases", Feldt and Brennan offer a reminder that this measurement is set in the behavioral sciences rather than the physical sciences (1989, p. 106). The measurement process itself is likely to effect a change in the examinee. A second concern is voiced about the specification of instruments that would provide such a true score. Feldt and Brennan argue that it must be possible to "define what meant is by interchangeable test forms without the use of the concept of true score, lest the definition be circular" (1989, p. 107). The third concern expressed by Feldt and Brennan is that conditions may vary legitimately from score to score, and the effects of subtle differences in obtaining scores might lead to redefinitions of true score. Such variations can lead to multiple reliabilities, even within a set population.

The shadow of "true score" is "measurement error", an observed score is made up of these two components. Feldt and Brennan identify three possible sources of error variance: (1) random variations within individuals (e.g., health, motivation, carelessness, luck); (2) situational factors such as the working environment of the examinee (factors may be psychological or physical); and (3) instrumental variables such as variations in

machinery, or effects arising from a misfit of the domain sampled for individual examinees, advantaging some and disadvantaging others. Feldt and Brennan describe the five assumptions of classical reliability theory:

1. An observed score on a test or measurement is the sum of a true score component and a measurement error component.

2. Parallel forms are measures constructed to the same specifications, give rise to identical distributions of observed scores for any very large (infinite) population of examinees, covary with each other, and covary equally with any measure that is not one of the parallel forms.

3. If it were possible for an individual to be measured many times, using different parallel forms on each occasion, the average of the resultant errors of measurement would approach zero as the number of measurement increased.

4. If any infinite population of examinees were tested via any given form of a test, the average of the resultant errors would equal zero, provided the examinees were not chosen on the basis of the magnitude of the observed test score.

5. Consistent with the foregoing assumptions, the true score of an individual is perceived as a personal constant that does not change from form to form. For convenience, the true score is regarded as the average score the individual would obtain if tested on an infinite number of parallel forms. (1989, p. 108)

Parallel forms are an important part of classical reliability theory. Feldt and Brennan consider the development of a large number of parallel forms as impractical. Instead they describe strategies for the use of two interchangeable test forms, and for the re-administration of a form previously used. The parallel-forms approach is considered the ideal, but administratively most difficult approach. Its main disadvantage stems from the reluctance of school authorities to retest students; in practice, available second forms tend to be ignored. For the test-retest approach, Feldt and Brennan believe that:

> A second administration of the same tasks or stimuli is a natural and appropriate procedure when two conditions are met:
> a) no significant changes in examinee proficiency or character is to be expected within the period of measurement, and
> b) memory of the tasks or stimuli will not influence responses on subsequent presentations. (1989, p. 110)

They warn that the second characteristic is unlikely to be met in pencil-and-paper administrations, as many examinees believe that they are being assessed on their consistency, and attempt to respond as they did in the first administration. This causes a falsely high level of reliability.

The estimation of reliability from a single administration of a test is based upon part-test similarity. In this approach the test itself is divided into two forms. In an ideal situation the mean and variance must be the same for each of the forms. As this is an extremely harsh condition, the measurement community has found ways of accommodating the reality of non-parallel forms, and alternative models have been created. Each of these models represents a progressive reduction in adherence to the ideal. Feldt and Brennan, (1989, pp. 110-111) describe these models:

- *tau-equivalent* forms have identical means, but differences in error variance and hence, in observed score variances;
- *essentially tau equivalent* forms exhibit mean differences in addition to variance differences, but the difference of the true score will be constant between parts for every examinee;
- *congeric forms* are even less parallel, but the true scores between parts are perfectly correlated in a linear sense, and have different error variances and score variances;
- *multi-factor congeric* forms attempt to identify systematic components within the forms and apply constants to weight each of these factors. The true scores of the two parts are represented by the same factors, in differing combinations.

These different types of form allow the calculation of various coefficients of reliability[17]. This discussion of reliability is focused upon the underlying assumptions of the theories rather than the modes of calculating coefficients, particularly as Traub and Rowley (1991) stress that:

> reliability is not simply a function of the test. It is an indicator of the quality of a set of test scores; hence reliability is dependent on characteristics of the group of examinees who take the test, in addition to being dependent on the characteristics of the test and the test administration. (p. 41)

Traub and Rowley assert that "inconsistent measurements are a bane to persons engaged in research" (1991, p. 37). While this statement may be at the core in consideration of closed assessment procedures where a single "right answer" is valued, there are many examples of performance assessment in science where students are asked to plan, design or experiment, with a range of possible outcomes or answers.

Generalizability Theory

Generalizability theory (G theory) has been used in analysis of student test scores in performance assessments by several researchers (e.g., Shavelson et al., 1992; Koretz, Stecher, Klein, McCaffrey, & Deibert, 1992; Candell and Ercikan, 1992). As performance assessments become used more widely, this approach to analysis of sources of variation in performance is likely to see greater use. Generalizability theory was first proposed by Cronbach, Gleser, Nanda, and Rajaratnam in *The Dependability of Behavioral Measurement* (1972). Much of this summary of G theory is taken from *Generalizability Theory: A Primer* by Shavelson and Webb (1991). Shavelson and Webb characterize G theory as "a statistical theory about the dependability of behavioral measurements" and within this definition they refer to dependability as:

---

17 Spearman-Brown, Cronbach alpha and Guttman among others.

> the accuracy of generalizing from a person's observed test score on a test or other measure (e.g., behavior observation, opinion survey) to the average score that the person would have achieved under all possible conditions that the test user would be equally willing to accept. (1991, p. 1)

A necessary condition, and limiting assumption, is that a person's ability or measured attribute remains in a steady state over different occasions of measurement. Thus differences among scores earned by an individual arise because of one or more sources of error in measurement[18]. Other methods of estimating error have generally been based upon classical test theory and are limited to estimates of one source of error at a time. The alleged advantage of G theory is that it is possible to estimate the value of each of several sources of error in a single analysis. Shavelson and Webb claim that:

> G theory enables the decision maker to determine how many occasions, test forms, *and* administrators are needed to provide a dependable score. In the process, G theory provides a summary coefficient reflecting the level of dependability, a generalizability coefficient that is analogous to classical test theory's reliability coefficient. (1991, p. 2)

Thus, by running a generalizability theory analysis (as done by Baxter et al., 1992; Shavelson et al., 1993; Shavelson et al., 1994), and by making several assumptions as to the impact of the types of error, it is possible to propose model studies in which the error is reduced to an acceptable minimum[19], comparable to those established for classical theory. A consistent finding in the reported results of Shavelson and his associates is that "measurement error is introduced by task-sampling variability, and not by variability due to other measurement facets" (Shavelson et al., 1993, p. 229). Shavelson, Baxter and Gao explain the consequences of this finding for performance tasks in large scale assessments:

---

18 "Error in measurement" changes a student's score from the "true score".

19 Shavelson and his collaborators have chosen to use a G value of 0.8 as the critical value about which to manipulate the number of tasks, studies or raters to find an acceptable level of dependability.

> Regardless of the subject matter (mathematics or science) or the level of analysis (individual or school), large numbers of tasks are needed to get a generalizable measure of achievement. One practical implication of these findings is that – assuming 15 minutes per CAP task, for example – a total of 2.5 hours testing time would be needed to obtain a generalizable measure (0.80) of student achievement. (1993, p. 229)

In a later paper, Shavelson, Gao, and Baxter (1994) demonstrate the influence of task domain upon task sampling variability. Shavelson, Gao, and Baxter revisit some of the group's earlier work (Shavelson, Baxter, Pine, Yur, Goldman, & Smith, 1991) in which students investigated the preferences of sow bugs for different environments: (a) light vs. dark conditions, (b) damp vs. dry conditions, and (c) factorial combinations of these environments (light and damp, etc.). Shavelson, Gao, and Baxter report that "Observations and subsequent data analyses showed convincingly that the third experiment was qualitatively different, and more difficult, than the first two" (1994, p. 3). In this 1994 paper, Shavelson, Gao, and Baxter base their analysis upon the premise that the content of the domain of the questions appropriately represents the domain or universe to which the researcher wishes to generalize. Shavelson, Gao, and Baxter argue that the factorial experiment is beyond the domain of fifth-grade science and as such represents part of an inaccurately specified domain for the set of tasks. However, if only the first two tasks are considered, then the domain is deemed appropriate. Shavelson, Gao, and Baxter consider this a significant development:

> The practical implications of this misspecification are telling. To obtain a coefficient of 0.80, 1 rater and 3 experiments would be needed for both relative and absolute decisions, when the domain is correctly specified. When misspecified, 1 rater and no less than 7 experiments would be needed for relative decisions, and 1 rater and 11 experiments for absolute decisions. (1994, p. 8)

With this "elegant" solution of eliminating the "inappropriate" experiment Shavelson, Gao, and Baxter found a way to enhance the generalizability of the two experiments. Shavelson and his colleagues claim that they can generalize performance on these two sow bug tasks to the domain of "living things" from the California Science Framework (1990). There are problems with content representation that make this claim troublesome for science

educators, who might argue that an acceptable representation of the Framework requires more than two similar tasks to control experimental variables for insects. An analysis of the Framework would likely show that content areas such as human biology and plant biology require some representation. Statistical manoeuvers such as those performed by Shavelson, Gao, and Baxter highlight the need for vigilance in the application of generalizability theory, and provide strong evidence to support Hein's (1990) plea for the involvement of science educators in the design and interpretation of science assessments.

The factors within a test that contribute to the reliability of test scores, test length, item type and item quality (Traub and Rowley, 1991) also apply in the application of generalizability theory to the analysis of the variance of test responses. Longer tests are considered more reliable and to provide a more generalizable set of scores, but the principal condition for the additional items is that they should:

> function in the same way as those already present. They should be of the same type (multiple-choice, short answer, etc.) and should test similar knowledge and skills. But also it is necessary that students should approach them similarly; if the length of the test is such that fatigue, boredom, or resentment begin to affect the students' behavior, we could not expect the Spearman-Brown formula to give us sensible predictions. (Traub and Rowley, 1991, p. 43)

The type of item, and its method of scoring, favours a large number of shorter, objective items. This generally eliminates scorer inconsistency as a source of measurement error and extends the represented content, reducing unreliability resulting from chance in question selection.

The issue of item quality deserves careful examination. Traub and Rowley (1991) warn that unclear or ambiguous items will lead to multiple interpretations by students, clearly reducing reliability. If items are too difficult then students will guess and introduce randomness to the scores; items that are too easy will be answered correctly by everyone and not distinguish between students. Traub and Rowley explain the issue clearly:

> Items that contribute most to test reliability are those that discriminate – in the technical sense, this refers to items on which students who possess the knowledge and skill needed to answer the question correctly have a better chance of success than students not in possession of this knowledge and skill. Items that are either very easy or very difficult for all the students being tested cannot be good discriminators. Therefore, it can be said that in order to maximize reliability a test should be pitched at a level of difficulty that matches the abilities of the students, neither too easy for them nor (the worse of the two) too difficult. (1991, p. 43)

These three factors pose particular problems for criterion-referenced tests. As most performance items have been designed with reference to specific criteria, the use of such reliability coefficients is necessarily inappropriate.

Similar concerns exist in examining the variance requirements of G theory. The set of tasks must enable the task sampling variance to be low (Shavelson et al., 1993). This means that students should perform at similar levels from one task to the next. By making the tasks very similar this can be achieved (Shavelson et al., 1994), but only at the expense of a reduction in content representation.

Instrumental Variables and Performance Assessment

In the context of complex performance tasks it is most appropriate to consider reliability in terms of consistency of the data collection procedures and the reproducibility of the data obtained. These procedures can be considered in terms of:

1. equipment stability;
2. administration procedures; and
3. uniformity in the data analysis procedures, leading to the application of consistent evaluative standards.

Stability of equipment used in hands-on assessment in science is vital. The equipment must not only perform to a well-defined specification, but must behave in the same manner from one student to the next, and in many different assessment sites. The administration procedures developed for an assessment must enable each student to have an equal

opportunity to perform at her/his best, but not enable any group of students to have a particular advantage. The training of the administrators must ensure that issues relating to fairness are clarified, and that procedures such as orientation time on tasks, and response to student questions are considered.

The issue of reliability of the rating of student performance has received considerable attention in the literature (Raymond and Houston, 1990; Slater and Ryan, 1993). Complex tasks which lead to qualitatively different responses can challenge the reliability of score interpretation, as for example in the Vermont portfolios (Koretz et al., 1992). However, Shavelson, Gao, and Baxter report more positive results for performance assessments in science:

> The findings are consistent. Inter-rater reliability is not a problem. Raters can be trained to score performance reliably in real time or from surrogates such as notebooks. (1994, p. 1)

Hardy (1992) reports a study in which inter-rater reliability coefficients ranged from 0.76 on one task to 1.00 on two other tasks. Both analytical and holistic scoring techniques were used to produce inter-rater reliability values of 1.00, with student samples of 183 and 240 respectively.

CHAPTER 3 — THE 1991 BRITISH COLUMBIA SCIENCE ASSESSMENT

AN HISTORICAL PERSPECTIVE

The 1991 B.C. Science Assessment is the fourth in a series of science assessments organized as part of the Provincial Learning Assessment Program (P.L.A.P.). The P.L.A.P. was set up in the 1970's by the provincial government to evaluate the effectiveness of the core program for education in British Columbia. The first science assessment took place in 1978 (Hobbs et al., 1980). Further assessments followed in 1982 (Taylor et al., 1982), and in 1986 (Bateson et al., 1986). These first three assessments were similar in structure, with three constituents:

1. Achievement tests for all students in Grades 4, 8 and 12 (Grade 10 in 1986);
2. A two-part instrument to survey students' attitudes towards science and scientists, and also to examine students' perceptions of the type of science teaching to which they had been exposed; and
3. Teaching/learning surveys for a sample of teachers in elementary, junior-secondary and senior-secondary schools.

These three parts provided sufficient data to enable the evaluators to make the wide range of inferences about the state of science education in B.C. demanded by the statement of major purposes of the P.L.A.P. These purposes are:

1. To monitor student learning over time;
2. To inform the public of the strengths and weaknesses of the public school system;
3. To provide the province and individual districts with information that can be used to identify strengths and overcome identified weaknesses;
4. To assist curriculum developers at the provincial and local levels in the process of improving curriculum and developing resource materials;

5. To provide directions for change in teacher education and professional development;
6. To provide information that can be used in the allocation of resources at the provincial and district level; and
7. To provide directions for educational research.

(British Columbia Department of Education, 1975)

The initial assessment in 1978 provided baseline data that have been used and extended by subsequent assessments to identify trends in student achievement. Anchor items from 1978, 1982 and 1986 were used in the 1991 assessment.

GOVERNANCE AND STRUCTURE

The management structure for the 1991 science assessment followed a model developed for earlier assessments. Approximately two years before the expected date of an assessment, the Ministry of Education requests proposals to conduct the assessment from teams of evaluators. Before proposals are submitted, Ministry personnel brief potential contractors about possible assessment instruments, sampling procedures, and levels of reporting. The proposals are reviewed by the Ministry and the contract for the assessment is awarded to a "Contract Team". Contract Teams have generally been based in university faculties of education, e.g., the University of British Columbia (U.B.C.) in 1978, the University of Victoria in 1982. A single coordinated proposal with a range of optional components was submitted for the 1991 assessment, and included personnel from each of the three universities in B.C.

Contract Teams are independent of the Ministry, but report to a Ministry-led management or review committee. This committee is made up of Ministry personnel, teachers of different grades, university/college instructors, parents, school trustees and members of the public. The Contract Team, in consultation with the review committee, prepares the materials for the assessment and presents the items and procedural details to

sub-committees of the management committee. These sub-committees review the items focusing upon the suitability of wording and content, and make recommendations for change. As the first three science assessments were all machine-scored, interpretation panels with a similar structure to the review committees met to examine student results. The range of components in the 1991 assessment led to significantly broader approaches. These are discussed below.

The 1991 B.C. Science Assessment saw a transition in the definition of the purposes of the assessment by the Ministry of Education. The report of the Royal Commission on Education Commission (Sullivan, 1988) and the subsequent Ministry response in *Year 2000: A Framework for Learning* (Ministry of Education, 1990) provided the impetus for the Ministry to seek greater breadth in the 1991 assessment. The emphasis on learner-focused curriculum and assessment in the *Year 2000* document led to a restatement of the objectives for the science assessment. These are to:

- Describe to professionals and the public what children of various ages CAN DO in the curricular area of Science
- Examine CHANGES in student performance over time, and provide baseline data for the proposed changes in educational programs
- Describe RELATIONSHIPS among instructional activities, use of materials, teacher and student background variables, and student performance and attitudes
- Describe the differences in student performance and attitude that are related to GENDER
- Provide EXEMPLARS OF ASSESSMENT TOOLS that can be used by teachers, schools, and districts for assessment of science processes and performances
- Provide directions for educational RESEARCH
- Suggest areas of need, and provide direction for decisions regarding both in-service and pre-service TEACHER EDUCATION. (Emphasis in original. Erickson et al., 1992, p. 1)

While the pedigree of these objectives can be traced back to the original set of major purposes, there are many significant changes in emphasis. The overall tenor of the

assessment objectives has become more focused upon students, teachers and the practices of education, and appears to reflect the outlook that education is about people and what they do together. These revised objectives are much more tentative in tone, and indicate that there has been a recognition that changes in education cannot be forced from above. For example, "identify" becomes "suggest" when dealing with teacher-education. Equal opportunity has become an issue in science and assessment, so there is an "up front" requirement to identify and describe gender-related differences.

The first objective in the list is pertinent to the discussion of performance assessment. The change from "the strengths and weaknesses" to "what children of various ages can do" is evidence of a philosophical metamorphosis. This "kinder gentler" approach towards assessment led to the development of the four components of the assessment. The relationship between the four components is shown in the Overview (Figure 4, next page).

Component 1: The Classical Component

This component retained many of the features of the earlier assessments – multiple-choice achievement items, student background survey and attitude scales, and a teacher questionnaire. An extension to the procedure was made through the use of open-ended items. The Ministry chose to report results at the provincial rather than district or school level. This led to the choice to use a 30% sample[1] rather than blanket coverage of all students in Grades 4, 7 and 10. In addition, data were collected from 10% samples of students in Grades 3, 5, 6, 8 and 9.

---

1 The sample was chosen to represent the Province by geographic zones, and also by school size.

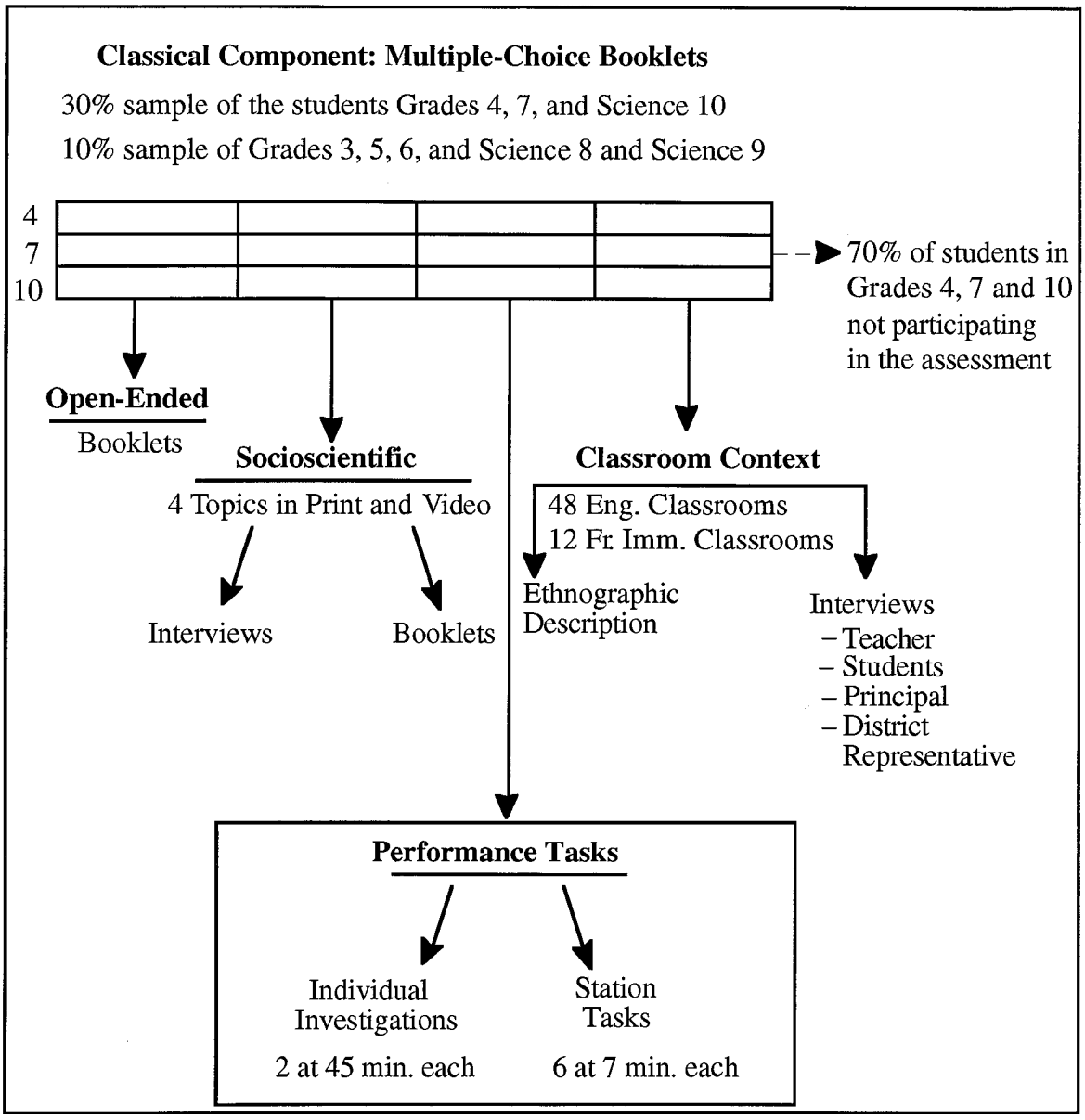


Figure 4. Overview of Science Assessment Components
(taken from Erickson et al., 1992, p. 5)

Component 2: The Student Performance Component

In this component of the assessment the Contract Team identified and developed two distinct modes of assessment, "stations" and "investigations", both of which involve student interaction with materials. Students in Grades 4, 7 and 10 participated in this component, with equal numbers working through each mode. For stations and

investigations, the results are reported in terms of descriptions of what the students actually did, and as well in terms of levels of student performance on a five-point scale.

Component 3: The Socioscientific Issues Component

> The socioscientific issues component was developed to:
>
> study student understandings of, or points of view with respect to, science, technology and society issues, and
>
> compare ways of assessing understandings of, or points of view with respect to, socioscientific issues. (Gaskell, Fleming, Fountain, & Ojelel, 1992, p. 3)

The Contract Team prepared videotapes for students in Grades 7 and 10. These present conflicting perspectives on four issues, e.g., clear cut logging, and the use of animals in scientific research. Through a series of interviews and open-ended questions, specific multiple-choice and written response instruments for each scenario were developed. Distributions of students' points of view, by gender, by grade and by medium (print or video) are reported.

Component 4: The Context for Science Component

The focus of this component was to provide "more information about classroom practices and the context within which students gain their knowledge and understanding about science" (Wideen, Mackinnon, O'Shea, Wild, Shapson, Day, Pye, Moon, Cusack, Chin, & Pye, 1992, p. 9). The Contract Team chose to collect data from structured observation of 60 classrooms, and interviews with teachers, samples of students, principals and district representatives. In the results, the Contract Team describes patterns of classroom practice and discusses the influence of district support upon these practices. The report *British Columbia Assessment of Science 1991 Technical Report IV: Context for*

Science Component concludes by presenting an "agenda for improvement" (Wideen et al., 1992, pp. 129-141).

STUDENT PERFORMANCE COMPONENT: A DETAILED DESCRIPTION

The time-frame of this component of the assessment spanned three years from the fall of 1989 until the Technical Report was signed off by the Ministry of Education and Ministry Responsible for Multiculturalism and Human Rights at the end of 1992.

Planning for the Assessment

Task development began with a review of the literature on performance assessment projects from around the world. The Assessment of Performance Unit (APU) for Science in England (Johnson, 1989) and the National Assessment of Educational Progress (NAEP) document *A Pilot Study of Higher Order Thinking Skills Assessment Techniques in Science and Mathematics* (Blumberg et al., 1986) were the prime sources of information about development and exemplars of potential tasks. In addition, the assessment programs that had been developed in New York (New York State Department of Education, 1989), California (Anderson, 1990; Comfort, 1990), England (CHASSIS edited by Wilson, 1986), Manitoba (1988), and Scotland (*Techniques for the Assessment of Practical Skills in Foundation Science – "TAPS"* by Bryce, McCall, MacGregor, Robertson, & Weston, 1984) were scrutinized. These assessments illuminated the need for breadth in the modes of assessment, particularly in representing content and a range of activities for the students. The NAEP pilot study of 1986 was heavily influenced by the work of the APU, and many of the features of these two earlier assessments were considered carefully in the planning and development of the Student Performance Component of the 1991 B.C. Science Assessment.

The Assessment Framework

The assessment framework is described in *The Assessment of Students' Practical Work in Science* (Erickson, 1990) a paper produced for the Ministry of Education at the time the Contract Team started preparations for the assessment. In addition to recognizing the influence of educational initiatives in British Columbia, Erickson identifies two other contexts that he considers significant in the development of an assessment framework: the context of the constructivist perspective and the context of assessment. Given the influence of the constructivist perspective, Erickson emphasizes the importance of experimentation:

> If we take a constructivist perspective of learning seriously then we see that a fundamental aspect of learning anything new consists of a form of continual experimentation. Thus at a very general level of description we obtain an image of children as well as adults constantly engaged in a process of constructing conjectures about the nature of the social and physical worlds in which they inhabit and testing these against "the reality" of those worlds. (Erickson, 1990)

That the framework, with its concentration upon the notion of experimentation, differs from those previously identified in curricular documents in British Columbia is seen by Erickson as a necessary and desirable consequence of this constructivist perspective.

Earlier science assessments had shown that few teachers engaged in assessing their students in practical mode; indeed as students progress through their education, the amount of hands-on science decreases (Bateson et al., 1986). It became an explicit focus of the Contract Team to exemplify experimentation to teachers:

> The specification of a framework of objectives in the area of students' practical work in science should provide teachers with a better understanding of the importance of practical work in science instruction and enable them to construct effective assessment strategies. (Erickson, 1990)

The framework is also intended to assist teachers in working towards assessing "their students' progress in developing the types of skills and competencies that they consider to be important in this particular curricular domain, considering the age and experience of the

students they are currently teaching" (Erickson, 1990). The language of the framework is deliberately general, a position that is justified by the statement that

> the framework is intended to be used as a basis for generating specific outcomes suitable for the developmental levels of the students to be assessed and embedded in an appropriate content area for that group of students – for instance, those content areas specified by the curriculum guide. (Erickson, 1990)

There are two levels of description within the framework. General level descriptors are called "dimensions". For example, "measurement", the "use of apparatus" and "communication" are classified as dimensions; the detailed levels of description, the sub-categories within each dimension, are called "abilities". The dimensions are shown in Figure 5; the complete table of dimensions and abilities is given in Figure 29 (page 152).

(1) Observation and Classification of Experience

(2) Measurement

(3) Use of Apparatus or Equipment

(4) Communication

- i) Receiving and interpreting information
- ii) Reporting information

(5) Planning Experiments

(6) Performing Experiments

Figure 5. Dimensions of Science

Framework descriptors require elaboration to include a specific context and content for each task. The Contract Team called these statements of specification "objectives". For example, the station "Rolling Bottles" is described using the following statements:

1. Students observe three bottles filled with different amounts of sand roll down the ramp, and identify the fastest. 1.a
2. Students explain why one bottle went fastest. 4.b
3. Students observe an empty bottle roll down the ramp and compare its speed with one bottle previously rolled. 1.a
4. Students explain why one of the two bottles went faster. 4.b

(Erickson et al., 1992, p. 61)

It can be argued that to use only abilities 1.a (ability to describe observations made using the senses about a variety of living and non-living objects) and 4.b (ability to draw inferences from data presented in tabular, pictorial or graphic format or generated experimentally) to describe the station is insufficient. These "behavioral objectives" serve only to emphasize the salient features of each of the stations and act as guidelines for the evaluation of students' performance.

Stations and Investigations

The literature review identified many different modes of performance assessment: short exercises of two to six minutes in the TAPS scheme (Bryce et al., 1984), longer stations of 13 to 15 minutes in the California Assessment Program (Anderson, 1990), and investigations of up to one hour in the APU (Johnson, 1989). In examining the tasks through the lens provided by the B.C. assessment framework, it was perceived that the shorter tasks tend to be prescriptive in nature, with students following instructions and responding to short questions; the majority of these tasks fit into Dimensions 1 to 4. Investigations are open-ended tasks in which students design and perform experiments to solve specific questions; such investigations appear to focus upon Dimensions 5 and 6.

A limit of one hour of contact per student was placed upon the assessment by the Ministry of Education. The Contract Team chose to use two distinct types of task to broaden the range of the assessment in an attempt to cover the major part of the spectrum of school science. The two distinct types of assessment task are:

Stations: in which students spend seven minutes working through short, time-limited tasks. Each station focuses upon different dimensions and abilities in diverse contexts of school science. Students work through one of two different circuits, each comprised of six distinct stations. The data for station tasks are student's written responses.

Investigations: in which students work in pairs on a problem, presented as an operational question, and use a specific set of materials. Each pair of students spends between 40 and 50 minutes upon a single investigation. The data for investigation tasks consist of observers' records, students' written responses, and students' verbal responses in an interview following the investigation.

The process of pilot-testing is described in *British Columbia Assessment of Science 1991 Technical Report II: Student Performance Component* (Erickson et al., 1992).

> A variety of tasks used in the earlier mentioned assessment projects were pilot-tested for possible use in our station format. New tasks were created by members of the Contract Team to broaden the curriculum coverage and extend the range of dimensions and abilities assessed. During the cycle of pilot testing, attention was paid to student reactions to the questions within each task, and student criticism of each station was considered. Each of the stations was pilot-tested with at least 40 students from at least three different schools. (Erickson et al., 1992, p. 10)

Stations that were "borrowed" from other assessment jurisdictions were presented to the students for pilot-testing as they had been used in those regimes. The Grade 4 students had significant problems with the language of some instructions, and there was much deliberation, both with students and within the Contract Team, about refinements to the tasks. Further decisions were made by watching students perform the tasks in revising the structures of both instructions and the student responses sheets. Some assessments, for example the New York assessment at fourth grade (New York State Department of Education, 1989), separated the instruction sheets from the response sheets; others integrated the instruction sheet with the response sheet. The integrated sheet appeared to work better with the tasks used in B.C., particularly when the response sheet was limited to a single page. Twenty-two stations were eventually considered suitable for use in the

assessment. As four stations overlapped all three grades, and six stations overlapped two grades (either Grades 4 and 7 or Grades 7 and 10) it was possible to allocate 12 stations for each grade. This was done by constructing two circuits, each of six stations. Figure 6 shows the relationships between overlapping stations.

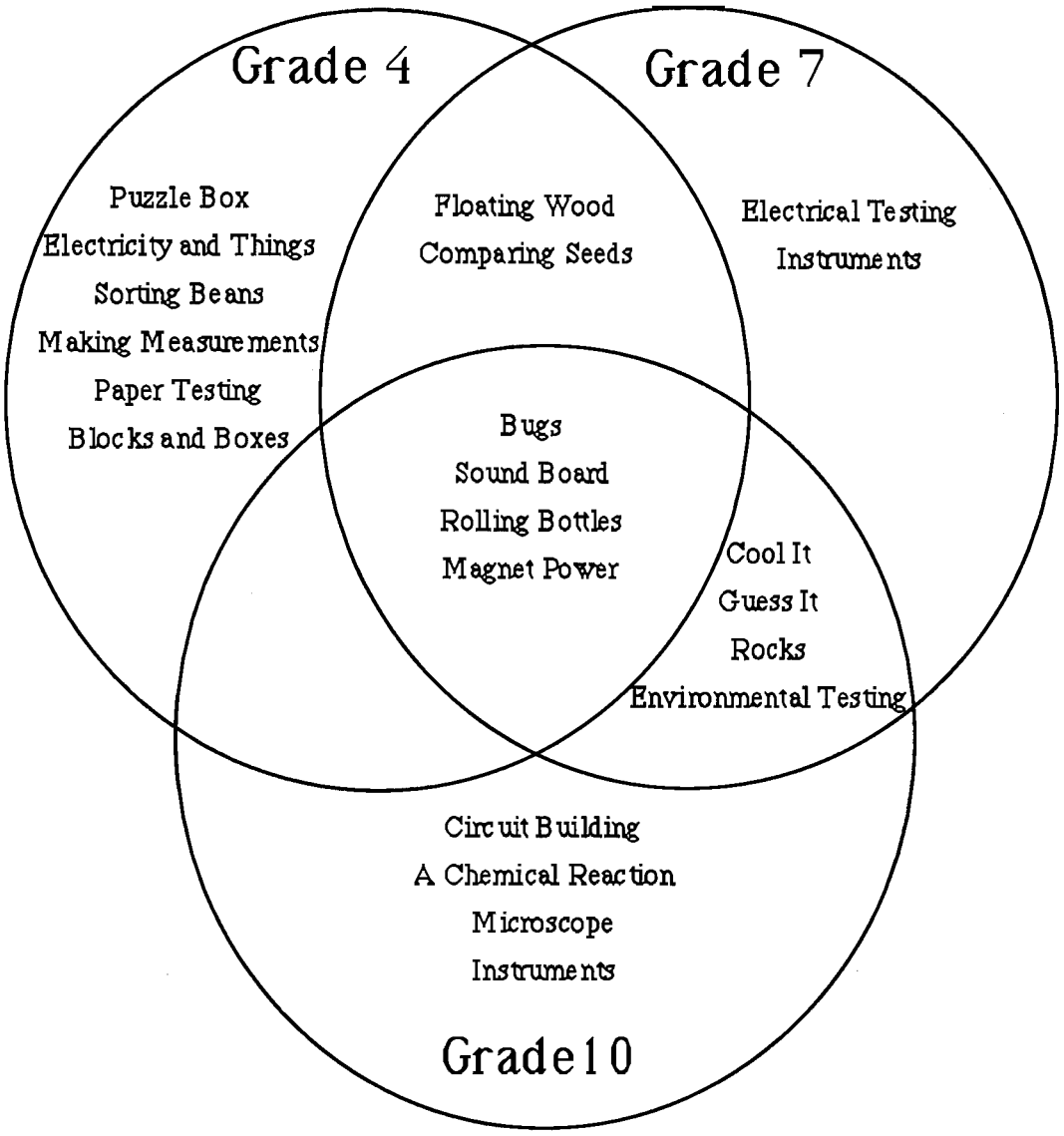


Figure 6. Venn Diagram of Stations
(Erickson et al., 1992, p. 11)

The pilot-testing of the investigations led to many decisions after each school visit. Two critical decisions were made about the structure of the investigations following the experiences with the pilot testing. The first of these was to assess students working in

pairs, rather than individually, as had been the procedure with the APU. This decision was made because students appeared to be more comfortable working with a partner, and it was easier for observers to identify procedures as students discussed their work. The second decision was that only two investigations were to be chosen for the assessment. "Magnets" is based upon Meyer's doctoral work (1992) and "Paper Towels" is derived from the work of the APU (Gott and Murphy, 1987). Other investigations were pilot-tested but rejected on conceptual or administrative grounds. The investigation "Survival", developed by the APU (Driver, Child, Gott, Head, Johnson, Worsley, & Wylie, 1984), involves students modeling a human body with an aluminum drink can which contained hot water. While many students visualized the model effectively, few were aware that the amount of time needed to cool the can was several minutes rather than seconds, an effect of the magnitude of the specific heat of water upon the rate of cooling. Administrative problems with live animals, snails and mealworms, led to the rejection of investigations based upon animal behaviour! The observation schedules for the investigations were developed as the pilot-testing proceeded and members of the Contract Team became more familiar with the experimental procedures. Although each investigation required its own observation schedule, the basic structure was maintained for both.

Figure 7 summarizes the cycle of pilot-testing for both stations and investigations, covering the phases of students working through the tasks, discussion of the tasks with members of the Contract Team through to revision of the tasks for further piloting.

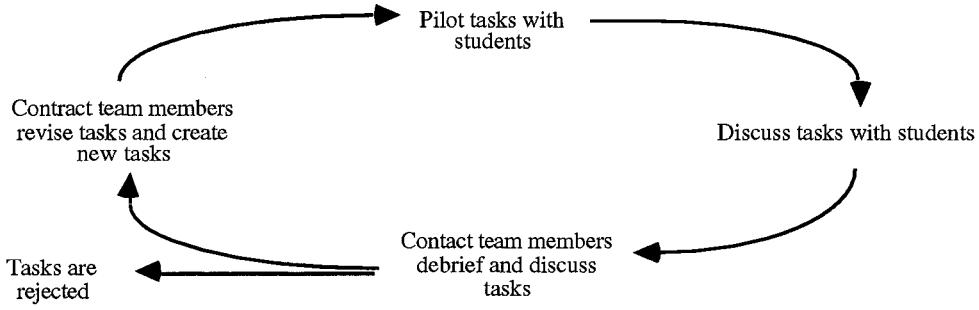


Figure 7. Cycle of Pilot-Testing

To direct the choice of particular stations and investigations for use in the assessment, the assessment goals were operationalized in terms of the questions shown in Figure 8. Four clusters of questions were used for both stations and investigations. The selection of tasks was complete before the third rotation of pilot-testing.

STATIONS	INVESTIGATIONS
Engagement	
Is the station interesting and likely to engage the student?	Is the investigation interesting and likely to engage the student?
Appropriate to student abilities	
Is the content knowledge appropriate for the grade level? Can the student use the equipment? Can the student complete the station in the allotted time? Does this station assess different abilities from the other stations?	Do students have the knowledge of the materials and procedures to answer the operational question? Do the operational question and the materials facilitate variability in student performance? Can students make appropriate measurement using the supplied equipment?
Appropriate task characteristics	
Will the materials stand up to repeated use? Will the equipment provide consistent results? Is the equipment available at reasonable cost? Can the equipment be transported easily?	Can an observer describe student performance, and reconstruct what the students did? Can students complete the investigation in the allotted time?
Range of appropriateness	
Is the station suitable for more than one grade level?	Is the investigation appropriate for all grade levels?

Figure 8. Criteria for the Choice of Assessment Tasks
(Developed from Erickson et al., 1992, p. 10 and p. 12)

Sampling

The overall structure of the assessment required that students who participated in the Alternative Components be a subset of the 30% sample in the Classical Component. Geographical representation from around the Province was an important element in the sampling design. The assessment coordinator divided the Province into six regions, each composed of approximately 12 school districts. Data were collected from a random sample

of three school districts in each region, with equal numbers of students in each district working through either stations or investigations – approximately 115 per type of task at each grade level. The unit of sampling was a complete class, and students in the class were allocated randomly to either stations or investigations by classroom teachers.

Teacher Preparation for Data Collection

Having identified grades and districts, the Ministry of Education funded two classroom teachers from each grade in each region (36 teachers in total) to collect and interpret the data. These teachers were nominated to the assessment by district administrators. The nominees attended an orientation workshop for data collection. This workshop took place at U.B.C. and included teachers working through both sets of tasks as though they were being assessed. After debriefing, the teachers administered a "dress rehearsal" of each mode of assessment to students from local schools. Members of the Contract Team took the teachers through the use of the observation schedules for both investigations, and helped use them with the "dress rehearsal" students. The orientation workshop concluded with the teachers checking the contents of the kits they were to use for data collection in their own regions.

The details for the tasks and the assessment procedures are presented in *Student Performance Tasks: Administration Manual* (Bartley, 1991). This manual contains a set of protocols for use by the teacher/administrators in introducing both investigations and stations to the students. In addition, equipment lists for all tasks are included, many with diagrams illustrating specific equipment.

Preparation for Data Analysis

The data from the station tasks are contained in the students' completed response booklets. These booklets were stamped with a unique identity number on each page. The

booklets were then taken apart and sets of station response sheets were created to enable coding to proceed with teachers not being aware of the identity, location or gender of the student. For the investigations there are multiple sources of data, including observation schedules, student response booklets, and notes from the teachers' interviews with students.

The Contract Team chose to develop coding sheets for both stations and investigations. The station coding sheets contain two parts, a descriptive part that was constructed using a random sample of 25% to 30% of student response sheets to generate response categories for each question, and an evaluative part which consists of one or two questions focused upon the significant aspects of each station. Generation of these "judgement questions" entailed extended discussions within the Contract Team. The format of the questions is consistent from station to station; the stem for all evaluations is "In your judgement, how well did the student...?". Judgements were made on a five-point scale where the central point is a "satisfactory" level of performance. The scale is shown as Figure 9.

1.	Not at all
2.	
3.	Satisfactory
4.	
5.	Extremely well

Figure 9. Scale for Evaluating Performance on Station Tasks – "In your opinion how well did the student...?"

For the investigations the Contract Team identified five questions that covered the salient features of student investigations. These questions are shown in Figure 10.

Evaluative Questions for Investigations

1. How well do the students plan an experiment to answer the question?
2. How well do the students develop a suitable measuring strategy?
3. How well do the students interpret the data collected to answer the question?
4. How well do the students report the results of their experiment?
5. In your judgement, considering ALL of the students' experiments how would you rate the quality of their performance on this task?

Figure 10. Evaluative Questions for Investigations

This holistic approach towards evaluating student performance required criterion descriptors for each level of performance. The Contract Team chose to leave the writing of the descriptors to the teachers at the coding workshop; it was believed that this would bring the specification of appropriate levels of performance, and attendant descriptors, closer to classroom practice.

Coding Workshop

The coding workshop took place over three days in July 1991. The teachers who had collected the data returned to complete the coding, spending equal time on stations and investigations. For the stations the teachers worked in teams of four; each team was responsible for coding four stations. Sets of response sheets for each station were prepared for teacher-orientation and development of criteria for levels of performance. These sheets were copied on brown paper, with four sets of five response sheets for each station. The coding sheets that were used at this stage also were printed on coloured paper. At the conclusion of coding a station, a further set of five response sheets, printed on red paper, was given to the teachers for collection of data on the consistency of their coding.

As most teachers had specialized in observing a specific investigation, coding of the investigations took place in two groups. Teachers coded the investigations performed by the student pairs that they had observed. The Contract Team asked the teachers to code a maximum of four different experiments[2] as some pairs of students had been identified as performing as many as 11 different experiments in the Magnets investigation! The observing teacher was required to identify the "best experiment", performance on which was evaluated using the questions shown in Figure 10. In addition, the teachers were asked to evaluate the students' overall performance in the set of experiments they had observed.

Analytical and Statistical Procedures

The coding sheets were passed on the Educational Measurement Research Group (EMRG) at U.B.C. Data from these sheets were entered into ASCII files for use with Statistical Package for the Social Sciences (SPSS) software. The basic analysis involves aggregation of the numbers and percentages of students coded for each descriptive category, and at each level of performance for the judgement questions. The results are presented in terms of the percentage of all students in that grade, and also percentages by gender. Thus, descriptions of what students did in performing the assessment tasks, and how well they performed are presented as percentages of the sample who participated in these tasks.

Further analysis of the data led to the production of correlation matrices of teacher judgements for each circuit of stations and also for the teacher judgements for each

---

2 A "different experiment" is one where the students appear to be changing their experimental approach. This might be by using different apparatus or by controlling different variables. Approaches and apparatus are shown on the coding sheets.

investigation. In addition, factor analysis tables were generated, but are not reported in the *Technical Report* due to time constraints. At the inception of the assessment it was hoped that data for the Student Performance Component of the 1991 B.C. Science Assessment could be merged with that from the Classical Component. Again, time constraints prevented this work from being completed.

Reporting and Interpretation of Results

The results for the assessment are reported in the *British Columbia Assessment of Science 1991 Technical Report II: Student Performance Component* (Erickson et al., 1992). Aggregated data for each station are followed by a discussion of the significance of the data for those students. Also included are the teacher-developed criteria for judgements at each station. Discussions for each station are broad ranging. They include an analysis of the descriptive data (focusing upon the nature of student approaches to the tasks) and also cover the evaluative data, with an analysis of the impact of the teacher-developed criteria upon the percentages of students performing at each level. Contract Team members specialized in writing about particular stations, and where tasks were used across grades, a single author prepared all the discussions. The comments conclude with an interpretation of the students' overall levels of performance on each station and some incorporate a critical evaluation of the virtues and deficits of the specific station.

Results from the investigations are examined in a single chapter in the assessment report. This chapter presents the data for all three grades, together with the interpretations made by the Contract Team. This arrangement facilitates comparisons between the experimental approaches of students in different grades.

The *British Columbia Assessment of Science 1991 Technical Report II: Student Performance Component* (Erickson et al., 1992) contains a chapter entitled "Issues in the Assessment of Student Performance on Science Tasks". Chosen for consideration are two

distinct facets: (1) comparative analyses of performance, and (2) specific technical issues concerning the validity of the assessment process. In comparing student performance in these assessment tasks the Contract Team chose to look at gender-related differences and grade-related differences. Both these issues had been identified by the Ministry of Education as needing to be addressed; these sections in the "issues" chapter are a response to the Ministry goals.

Gender-Related Differences

The section on gender-related differences in student performance shows tables of percentages of all students, female students, and male students who were evaluated as having performed at "satisfactory or better" levels, with differences between females and males of over 15% shown. The reason given for the selection of 15% as a notable difference is "pedagogical significance" (Erickson et al., 1992, p. 227). It was recognized that statistically significant differences are likely to be much smaller than 15%, but these would not translate easily into student numbers in a typical classroom of 20 to 30 students. Judgements scores with gender differences of at least 15% are shown in Table 1.

Table 1
Gender-related Differences in Student Performance (Derived from Erickson et al., 1992)

Grade	Station and Judgement	% Satisfactory or Better Rating		
		All students	Female	Male
4	4.7 Making Measurements: How well did the student measure temperature, length, volume and time?	91% (N=108)	99% (N=58)	82% (N=50)
4	4.10 Rolling Bottles: How well did the student explain the motion of the bottles (consider both "why" questions)?	63% (N=104)	71% (N=56)	54% (N=48)
7	7.7 Instruments: How well did the student measure selected properties, given a set of instruments?	46% (N=111)	34% (N=59)	58% (N=52)
7	7.11 Magnet Power: A. How well did student develop a strategy and use materials to identify the stronger of two magnets?	67% (N=111)	76% (N=59)	56% (N=52)
10	10.4 Environmental Testing: How well did the student draw inferences from collected data?	65% (N=106)	75% (N=44)	58% (N=62)

That only five out of the 57 judgements across the three grades showed differences of greater than 15% led the Contract Team to state that "similarities in performance between females and males were more evident than the differences" (Erickson et al., 1992, p. 223). Four out of these five differences appear to indicate superior performance of females, with variations from grade to grade.

The Contract Team did not report any pattern in gender-related differences in student performance for common stations. Measurement stations show a difference favouring females in "Making Measurements" at Grade 4, favouring males in "Instruments" at Grade 7, and "similar levels" in "Instruments" at Grade 10 (Erickson et al., 1992, p. 227). The station "Magnet Power" shows:

> at Grades 4 and 7 more females than males (differences of 14% and 20% respectively) were judged to perform at a "satisfactory or better" level in developing the strategy and using the materials; and
>
> at Grade 10 more males (74%) than females (64%) performed at a "satisfactory or better" level for the same aspect of the task.
>
> There was also a similar trend, but smaller differences, in the second teacher judgement on this station, communication of the strategy. This reversal of gender differences at Grade 10 is somewhat puzzling. Perhaps the open-ended nature of the problem and its solution appealed to the younger females (Grade 4 and 7) and showed in the quality of their responses, whereas the Grade 10 results may be related to students' experiences with the physics component of the junior secondary curriculum. At this age physics is often associated more with males' prior interests and knowledge (Erickson and Farkas, 1991) and so this may have created the different performance that we observed. (Erickson et al., 1992, p. 227)

The Contract Team is tentative in its interpretation of these differences, particularly in attempting to explain any pattern or cause. Other gender-related differences in student performance posed similar problems in explanation, and appeared to be quite small (less than 15%). Given earlier reported gender-related differences in performance in science assessments (Johnson, 1987; Robertson, 1987), the Contract Team was keen to restate that:

> gender-related differences are noteworthy by their absence in this study, especially in areas such as the physical sciences, where previous studies using more conventional modes of assessment have consistently reported large differences (e.g., electrical circuits). Many of the differences reported, whether showing males or females at higher performance levels, particularly with across grade variations are difficult to explain. Clearly, much more detailed work and analysis is required in some of these areas. (Erickson et al., 1992, pp. 228-9)

In concluding this section, the Contract Team comments upon the anecdotal data from teachers and students which indicate that students enjoyed working through the assessment tasks. Citing studies (Erickson and Farkas, 1992; Linn and Hyde, 1989) which emphasize the significance of "the influence of motivational and affective factors in producing gender differences on achievement tests" (Erickson et al., 1992, p. 229), the Contract Team speculates that this mode of assessment, and this set of tasks, have reduced gender-related differences. The Contract Team argues that this is an important step toward more "gender-fair assessment" (Bateson and Parsons, 1991).

Grade-Related Differences

The Contract Team chose to limit the analysis of grade-related differences in student performance to the four stations that were common to all three grades in the belief that "these four stations are sufficient to demonstrate and illustrate the grade-related issues of note in this type of assessment" (Erickson et al., 1992, p. 229). In choosing to make these specific comparisons the Contract Team made an effort to identify potential problems:

> The criteria for these judgements were made by the teachers who administered and coded these stations at that grade level. Hence for some of these common stations both the criteria and the way in which the criteria were interpreted differed significantly across the grade levels. For other, multi-grade stations there was more uniformity in these criteria. (Erickson et al., 1992, p. 229)

In working with these different criteria for teacher judgements, Erickson and his colleagues are necessarily cautious in making claims about differences in student performance. However, they express greater confidence in the comparisons of the descriptive data.

Pages 231 to 242 of the *British Columbia Assessment of Science 1991 Technical Report II: Student Performance Component* (Erickson et al., 1992) consist of tables of comparative data. The Contract Team reports many similarities in performance:

> For Sound Board the modal response category is the same for all three descriptive questions; for Rolling Bottles it is the same for two out of five questions; and for Magnets it is the same for three out of four questions. Likewise, if one creates a rank order in terms of the frequency of responses in these categories, a similar pattern is obtained from these descriptive questions. (Erickson et al., 1992, p. 243)

These findings were clearly not anticipated. The Contract Team makes two conjectures:

> The first is that the tasks did not require the kind of abilities and knowledge beyond the sort of "everyday knowledge" that most Grade 4 students have already constructed and hence the tasks did not discriminate in ways that we thought they might. A second conjecture is that the teachers implicitly interpreted the data in ways appropriate to the age of the students with whom they worked. (Erickson et al., 1992, p. 243)

These two hypotheses serve different purposes in explaining the findings. The first is seen as most suitable in accounting for the many similarities in student performance in the assessment tasks, particularly when the Grade 4 students demonstrate "considerable physical and linguistic experience with objects like stringed instruments, rolling objects, magnets, and insects in different stages of their life cycle" (Erickson et al., 1992, p. 243). Only where the task requires a more elaborate response is there evidence of substantive differences between the younger and older students. The student explanations in "Rolling Bottles" are used to exemplify this point, as older students showed a "greater degree of linguistic fluency than is available to most Grade 4 students" (Erickson et al., 1992, p. 243).

The second conjecture, the age-appropriateness of teacher interpretations of the data, most likely explains the nature of the criteria developed by teachers for the judgements. This is evident in that many of the criteria for higher ratings require students to provide extended responses or greater numbers of responses. The Contract Team considers that:

> these types of criteria are more difficult for Grade 4 students to meet because they generally lack the verbal fluency of the older students and they simply do not work and write as quickly. (Erickson et al., 1992, p. 243)

The data analysis and an extended discussion of the investigation results are presented in Chapter 6 of the *Technical Report* (Erickson et al., 1992). The Contract Team saw this section as an appropriate place for discussion of the question: "What is developing over the three grade levels as students are engaged in these types of open-ended investigations?" The results for the investigations parallel those for the stations in that there are many similarities in the performance of students in all three grade-levels. The Contract Team reports that:

> in both investigation tasks one is immediately struck with is the apparent *lack of differences between the students at the three age levels especially as it pertains to the planning and performance aspects of the tasks.* (Italics in original. Erickson et al., 1992, p. 243)

Detailed examination of experimental approaches for each investigation shows similar choices of both strategies and control of variables. The Contract Team interprets this as strong evidence in support of the use of complete investigations with younger students. The Contract Team also makes explicit that this evidence challenges the position of those favouring a step-up approach from "process skills" to investigations:

> What does appear to be changing for older students is:
>
> the construction of more elaborate and powerful explanatory models that are used to frame experiments of the type that the teachers observed in this project.
> (and)...
> we do see an increase in the abilities of students to perform adequately some experimental abilities such as conducting appropriate measurements with care and precision, identifying and controlling variables thought to be important to the outcome of the experiment, and finally providing an interpretation of the data and communicating that interpretation to others.
> (Erickson et al., 1992, p. 244)

Inter-coder Consistency Ratings for the Stations

The reliability of teacher coding and rating of student performance is a major issue in the use of performance assessments; indeed warnings about the use of holistic scoring systems have been made. Bryce and Robertson contend that:

> teachers should no longer be urged to judge or rate "holistically" their pupils' performances on "experiments" or "scientific investigations". In a comprehensive examination of the international literature we have recently shown that, however desirable, there is no demonstrable evidence of validity and reliability in currently available versions of assessment by holistic teacher-judgement. (1986, p. 63)

In the Student Performance Component of the 1991 B.C. Science Assessment, steps were taken to maximize the consistency of the teacher coding and evaluations. A structured set of procedures saw teachers working with practice sheets to develop the criteria for their judgements; this was followed by coding of actual student responses, and then checking consistency with yet another set of response sheets.

The Contract Team reports that for the stations the inter-coder consistency is acceptable (see for example Baxter et al., 1992), and in the range 0.83 to 0.86 for all codings and 0.62 to 0.81 for the teacher judgements[3]. Grade 4 teachers show greater levels of consistency and Grade 10 teachers lower levels; this difference is considered as:

> likely to have arisen from the widely different set of criteria each group of teachers had developed and the increase in complexity of the criteria used by the Grade 10 teachers. (Erickson et al., 1992, p. 248)

---

3 The term 'inter-coder consistency' refers both to descriptive and to evaluative codings completed by the teachers. When data are presented for consistency of the teacher judgements, these figures are synonymous with inter-rater reliability.

Coding consistency is considered an essential precursor to the subsequent issue of comparing student performance across tasks. The Contract Team believes that the consistency (reliability) is sufficient to proceed to this next step.

Student Performance Across Tasks

Although there are 12 stations for each grade, these were arranged into two circuits of six stations, with each student completing only one of the circuits. This arrangement of tasks enables comparisons of an individual student's performance within circuits A or B rather than between the circuits. As stated above, the Contract Team is confident enough in the reliability of the coding to discuss the issue of student performance across the tasks.

The consistency of student performance across stations is evaluated by the use of correlation matrices of individual student scores for each circuit. The Grade 7 matrices are discussed in the body of the Technical Report (Erickson et al., 1992) while the other matrices are published in the appendices to the report. The patterns for each of the six matrices are similar. These are:

1. Where there is a strong correlation, greater than 0.5 ($p < 0.001$), between student performances, it is usually between two judgements within the same station.
2. Correlations between stations with similar objectives or "skills" set in differing content areas are very low.

The Contract Team concludes that these data indicate that performance is influenced by a strong context effect "in the type of content knowledge and experience embedded within the task" (Erickson et al., 1992, p. 251); a conclusion that explains the within-station correlations. For the "between skills" correlations the Contract Team proposes that:

> These and other low correlations between judgements based on the assessment of similar "skills" in different stations suggest that the assessment of student performance may be more dependent upon the actual task context and less dependent upon the student possessing some generalizable "skill" or ability that can be applied equally well to a variety of tasks. (Erickson et al., 1992, p. 251)

Correlation matrices of the five teacher judgements in the investigations are seen by the Contract Team as "strongly supportive of the above claim that student performance appears to be very dependent upon the content understanding brought to these tasks by the students" (Erickson et al., 1992, p. 252).

Atomistic versus Holistic Scoring

The options for scoring student performance in hands-on assessment appear to range from atomistic or analytical at one extreme, to holistic at the other. The reliability of holistic scoring in science has been questioned by Bryce and Robertson (1986), but recent work by Baxter, Shavelson, Goldman, and Pine (1992) provides convincing evidence for the reliability of some holistic scoring methods. In explaining its decision to choose the holistic approach, the Contract Team writes:

> We have reviewed the literature, talked to the proponent of each perspective, solicited the views of local teachers and researchers, and examined our own proclivities and decided to be holistic. (Erickson et al., 1992, p. 252)

Holistic scoring enabled the Contract Team to ask the teachers to consider each station and make "a global judgement based on their professional experiences of students at that grade level" (Erickson et al., 1992, p. 252). In recognizing the professional judgements of the teachers in developing criteria for scoring rubrics, the Contract Team concedes that these criteria may not be viewed as acceptable by teachers from other school districts. The Contract Team concludes this section with a plea:

> We, at all costs, wanted to avoid the return in science teaching to the days of the Science as Process Approach where individual so-called "process skills" were taught separately. (Erickson et al., 1992, p. 253)

Project Recommendations

Although the *British Columbia Assessment of Science 1991 Technical Report II: Student Performance Component* (Erickson et al., 1992) contains the significant details of the procedures and the results, it is the chapter entitled *Performance Assessment Component* (Erickson, Carlisle, & Bartley, 1992) in the *British Columbia Assessment of Science: Provincial Report 1991* (Bateson et al., 1992) that presents detailed recommendations for science educators and policy-makers in British Columbia. The four recommendations that appear in the concluding section "Looking to the Future", are written in terms of actions that will lead to improved student performance in a range of contexts. These recommendations are:

1. Given the students' demonstrated abilities and interest in these assessment tasks, students should be given more opportunity to generate questions and to seek answers from their own investigations.

2. Given that students do not report results well, particular attention should be given to this aspect of investigations.

3. Given the positive student response to these tasks, we would encourage teachers to use more of these types of performance tasks in their assessment procedures.

4. Given the importance of content knowledge in the planning and execution of investigations, it is critical that the acquisition and use of relevant content knowledge be recognized and encouraged.

(Erickson, Carlisle, & Bartley, 1992, p. 39)

The first three recommendations are proposals for change in the practice of science teaching, while the fourth recommendation advocates a change in thinking about school science, and has significant implications for elementary school curricula.

Because the Student Performance Component of the 1991 B.C. Science Assessment was successful, a package entitled *Science Program Assessment Through Performance Assessment Tasks: A Guide for School Districts* (Bartley, Carlisle, & Erickson, 1993) was developed. Ministry personnel are employed in the dissemination of this resource around the Province, and several districts have developed plans for their own

assessment procedures. The package is self-contained and enables districts to collect and interpret their own data with the use of common microcomputer software.

CHAPTER 4 — A FRAMEWORK FOR THE VALIDATION OF THE USE OF PERFORMANCE ASSESSMENT IN SCIENCE

VALIDITY IN THE ASSESSMENT CONTEXT

In this chapter I describe and exemplify the essential components of a systematic framework for the development and administration of performance assessments in science. In my view, such a framework should attend to all significant decision-points right from the initial planning stages through to the analysis of both short and long term consequences of the assessment practice. I believe that employment of such a framework by an assessment developer, or user, is sufficient to validate the inferences and consequences of a specific assessment activity. The formal requirements for validation inquiry direct the focus towards an evaluation of the evidence and theoretical justifications to support the suitability of the inferences and actions based upon test scores (Messick, 1989b). I intend to substantiate the claim that the validation should extend from genesis to conclusion.

This chapter begins with an interpretation of current conceptualizations of validity (Messick, 1989a, 1989b; Moss, 1992; Shepard, 1993) which led to the identification of a set of criteria for the validation of performance assessments, previously described in Chapter 2 (Linn et al., 1991). As the criteria developed by Linn, Baker and Dunbar do not address all the issues that arise in the validation of performance assessment in science, I propose a more specific set of questions set in the context of school science. The use of these questions is exemplified by an evaluation of the validity of the inferences made in the *British Columbia Assessment of Science 1991 Technical Report II: Student Performance Component* (Erickson et al., 1992) and subsequent developments in science education in British Columbia.

The definition of Validity used in this study is that presented by Messick:

> validity is an **integrated** evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of **inferences** and **actions** based on test scores or other modes of assessment (Bold added. Messick, 1989b, p.13).

The focus upon an **integrated** judgement of inferences and consequential actions presents a sharp contrast to the traditional separation of validity evidence into content, criterion and construct. The association, or lack thereof, between these "types" is represented in Figure 11.

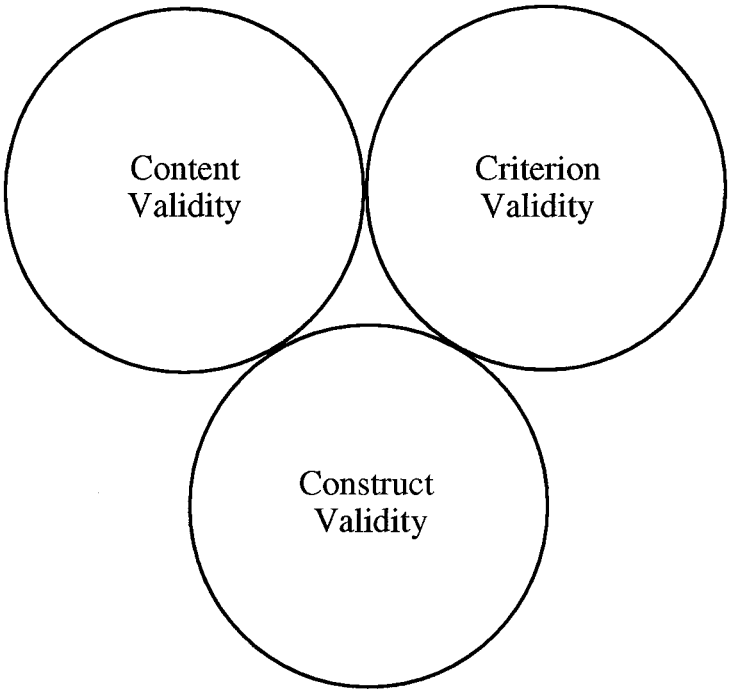


Figure 11. Traditional "Types" of Validity Evidence

Figure 12 illustrates an integrated view of validity based upon Messick's progressive matrix (1989a) where all "types" of validity evidence are subsumed under construct validity.

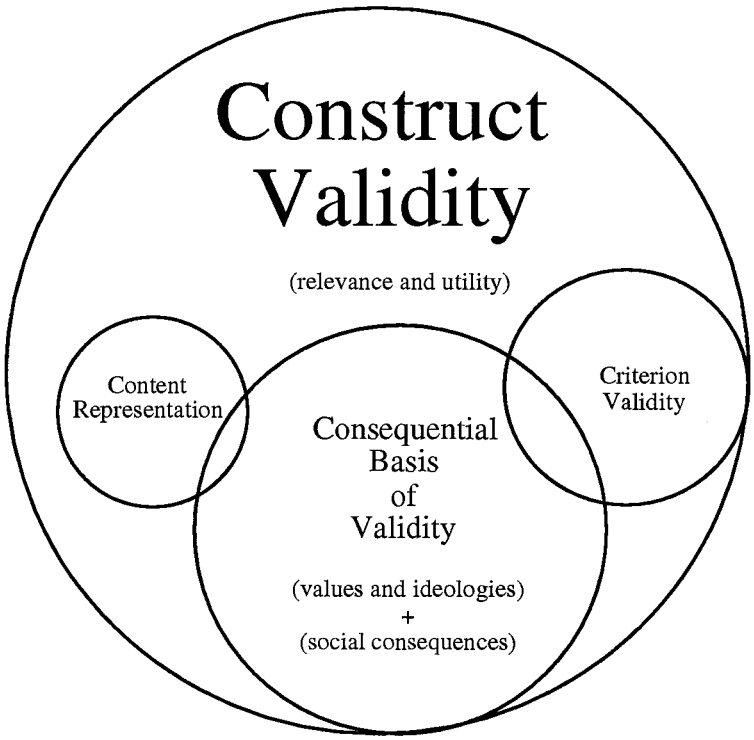


Figure 12. An Integrated Model of Validity Evidence

The significant sources of evidence for an examination of construct validity must include interpretations of score meaning, and also the consequences arising from the value implications of evaluators, teacher-administrators and students, as well as the social consequences of testing. Messick proposes that validation evidence should be collected to answer questions (1989b). Shepard identifies the central validity question as "What does the testing practice claim to do?" (1993, p. 429); other general questions, including "How well does the testing practice do what it claims to do?" and "What does the testing practice do beyond that which is claimed?" are implied. These general questions suggest which specific questions should be used for validation in a particular assessment context. In much the same way as the values of those involved in the management of a testing program lead to specific decisions and interpretations, the questions asked in a validation inquiry will manifest some value system.

The dominant values in the measurement community have been described as "the well established psychometric criteria for judging the technical adequacy of measures" (Linn et al., 1991, p. 16). These well established criteria include efficiency, reliability and comparability of assessment from year to year. Linn, Baker, and Dunbar speculate that such criteria would almost always favour the traditional multiple-choice assessment tasks in any comparison with the newer alternatives. They argue that as there is expansion in modes of assessment there must be expansion in the criteria used to judge the adequacy of assessments. In particular, Linn, Baker, and Dunbar caution that:

> Reliability has been too often overemphasized at the expense of validity; validity has itself been viewed too narrowly. (1991, p. 16)

Linn and his colleagues propose a set of validity criteria focused specifically upon performance assessments; these criteria relate to consequences, fairness, transfer and generalizability, cognitive complexity, content quality, content coverage, meaningfulness, and cost and efficiency. Moss (1992) finds these criteria to be within general standards (AERA, APA, & NCME, 1985). However, Messick (1994) exclaims that these criteria are more limited and "may become a problem for validation practice". In particular, Messick is concerned that this set of criteria might lead to insufficient emphasis upon "score interpretation and its value implications" (1994, p. 13).

A VALIDATION FRAMEWORK

The appraisal of the implications of underlying values is an important component of the understanding of validity in this study, and represents a significant expansion beyond the appraisal of score meaning. As the values of all the people involved in an assessment affect actions and inferences, the choice of questions in a framework for investigating validity must be examined as part of some value system. The questions can also be considered sequentially as they relate to concerns that should be examined at each stage of planning, administration, scoring, interpretation and recommendations, and subsequently

in reflection. The set of questions presented below meets the criteria that Linn, Baker, and Dunbar have proposed. In addition, it addresses the issues of score interpretation raised by Messick (1994). These eight focus questions were written with the specific purpose of guiding a validation inquiry of the Student Performance Component of the 1991 B.C. Science Assessment but are relevant to any hands-on assessment. They include:

(1) **Purposes of the Assessment**:
What are the explicit purposes of the assessment?
What are the operationalized purposes of the assessment?

(2) **Learning and Communication in Science**:
What models of science learning and communication are promoted by this mode of assessment?

(3) **Content Analysis**:
Are the assessment tasks appropriate for the students and within the curriculum?

(4) **Instrumental Stability**:
Does the equipment behave consistently over time?

(5) **Administration Stability**:
Are the administration procedures clearly developed and applied consistently?

(6) **Internal Consistency and Generalizability**
What factors affect the generalizability of student performance across tasks in a science assessment?

(7) **Fairness**
Do the assessment data indicate any bias towards or against any specific identified group?

(8) **Consequences**:
Were the intended consequences of the assessment achieved?
What were the unintended consequences?
What actions have been taken to support the "good" unintended consequences and abate the "bad" unintended consequences?

The analysis presented in this chapter explores the value implications of both these questions and the nature of responses to them.

Purposes of the Assessment

While it might appear that validation focuses solely on the after-effects of testing, the identification of purposes and intended consequences is a vital precursor in planning any assessment program. Shepard's question "What does the testing practice claim to do?" (1993, p. 429) reverberates through the inquiry. It is pertinent to ask "What does the testing practice **intend** to do?" as it is this question that guides the decision to assess, as well as the design of the assessment. Thus, in examining purpose, there appear to be two considerations, the "grand design", and the operational aspects of the assessment. This leads to two focus questions:

1. **What are the explicit purposes of the assessment?**
2. **What are the operationalized purposes of the assessment?**

The Ministry of Education of British Columbia funds and "owns" the provincial assessments. The Ministry identifies the purposes and modes of these assessments, but each actual assessment is conducted by an independent contract team which reports to the Ministry of Education. The Ministry of Education identified the following objectives for the 1991 B.C. Science Assessment:

- Describe to professionals and the public what children of various ages CAN DO in the curricular area of Science
- Examine CHANGES in student performance over time, and provide baseline data for the proposed changes in educational programs
- Describe RELATIONSHIPS among instructional activities, use of materials, teacher and student background variables, and student performance and attitudes
- Describe the differences in student performance and attitude that are related to GENDER
- Provide EXEMPLARS OF ASSESSMENT TOOLS that can be used by teachers, schools, and districts for assessment of science processes and performances
- Provide directions for educational RESEARCH

- Suggest areas of need, and provide direction for decisions regarding both in-service and pre-service TEACHER EDUCATION. (Emphasis in original. Erickson et al., 1992, p. 1)

The Contract Team for the Student Performance Component of the 1991 B.C. Science Assessment needed to operationalize these objectives. Erickson (1990) had explicitly identified the "constructivist perspective on learning" as the position from which the team would develop the assessment; this translated directly into the "CAN DO" objective receiving priority. Other objectives were also given major emphasis in the Contract Team's development of the assessment, particularly those related to "GENDER", "CHANGES" and "EXEMPLARS".

The assignment of high priority to the "CAN DO" objective had significant implications in the development of the tasks, the administration of the assessment, the coding and scoring of the data, and the interpretation of the results. Two types of hands-on task, "stations" and "investigations", were developed for the assessment. There was extensive piloting and discussion with students about the tasks, with many students reporting a personal sense of success in their own achievements. Tasks were designed with the intention of enabling all students to complete at least some part. A range of data collection procedures was developed and included teacher observations, student open-ended and structured pencil-and-paper responses, and interviews.

Those teachers who were to administer the assessment took part in a three-day orientation workshop. In this workshop they completed both stations and investigations under the proposed assessment conditions so that they would appreciate some of the complexities of the tasks, and be better able to assist students if/when equipment did not perform as required. The same teachers returned for a further workshop to code and evaluate the data describing student performance. Descriptions of student responses were entered onto coding sheets derived by the Contract Team from a sample of actual student response sheets. The Contract Team also identified one or two major facets of each station,

in terms of questions that the teachers should use to evaluate student performance. The common format for these questions was "In your judgement, how well did the student...?" Teachers were asked to rate the students on the five-point scale shown as Figure 9 (page 80). The criteria to describe student performance at each point on the scale were developed by examination of response sheets and discussion within grade-level groups of teachers.

The data were aggregated and are reported in percentages of students who made specific types of response. The evaluative questions are reported in tables, showing the percentages of students who achieved each level of performance. The Contract Team discusses these results in terms of the percentage of students who were evaluated as "satisfactory or better" (levels 3, 4 and 5) or "below satisfactory" (levels 1 and 2).

The technique for the description of student work in the investigations paralleled that for the stations, but used five questions by which teachers judged how well students planned their experiment, developed a measuring strategy, interpreted data, and reported data, with a final integrated judgement of students' overall performance.

An alternative interpretation of the "CAN DO" objective might have led to different task structures. For example, a set of tasks each with a single correct response would have lead to the results being reported in terms of Pass/Fail. Similarly, the tasks might have remained similar but analytical scoring could have been used with credit given for success in each major component of the task, an approach demonstrated in *Performance Assessment: An International Experiment* (Semple, 1992).

Other goals of the assessment were addressed as the assessment was planned, conducted and reported. Thus, tasks which covered a specific curriculum area considered essential, such as magnetism or electricity, were examined for gender bias, e.g., in familiarity with the equipment to be used, and changes were made to create tasks believed to be more gender-sensitive. Materials from a "female sphere of experience", e.g., paper

clips or hair pins, were used to balance other materials that were considered more likely to be encountered in the "male sphere of experience", e.g., steel washers. As this was the first assessment of this type in B.C., demonstration of change over time was not possible. The Contract Team chose to present similar tasks to students in two or more grades. It was hoped that this approach would enable the Contract Team to describe qualitative differences in student performance.

The elaboration of the purposes of an assessment is a multi-leveled process. While the "grand purposes" are passed down from the Ministry of Education and give direction to the assessment, the operational purposes can be elucidated only as the assessment is developed and negotiated between those involved. In the case of the Student Performance Component of the 1991 B.C. Science Assessment, the Contract Team, Ministry officials, teachers and students all participated in developing the working definitions of the purposes of the assessment.

Learning and Communication in Science

The influence of assessment upon instruction has been reported by many authors (e.g., Lovitts and Champagne, 1990; and Cole, 1991). Baron (1991) writes of tests as "magnets for instruction" and comments on the limiting effects of multiple-choice tests. Many of the arguments for performance assessment revolve around the differences in curriculum emphasis that comes from tasks in which students produce solutions rather than merely recognize or reproduce them (Newmann and Archbald, 1992). The focus question in this part of the validation inquiry examines the face validity of the tasks, i.e., the explicit message about science communicated to students and their teachers. The focus question for this section is:

What models of science learning and communication are promoted by this mode of assessment?

In his 1990 paper *The Assessment of Students' Practical Work in Science* Erickson identifies the constructivist perspective as central in many aspects of the assessment. Erickson also presents the framework which is used to describe student performance in the assessment. Six dimensions of student performance are proposed, each amplified by a set of more specific abilities. The dimensions identified by Erickson were shown in Figure 5 (page 72).

Stations, which typically occupy students for seven minutes, were examined by Contract Team members in an effort to identify the "dimensions" and "abilities" that are pertinent to the task. It was found that many of the stations are multi-dimensional, with an emphasis upon Dimensions 1 to 4. Not surprisingly, the investigations show an emphasis upon "planning experiments" (Dimension 5) and "performing experiments" (Dimension 6). Mapping of dimensions to type of task is shown in Figure 13

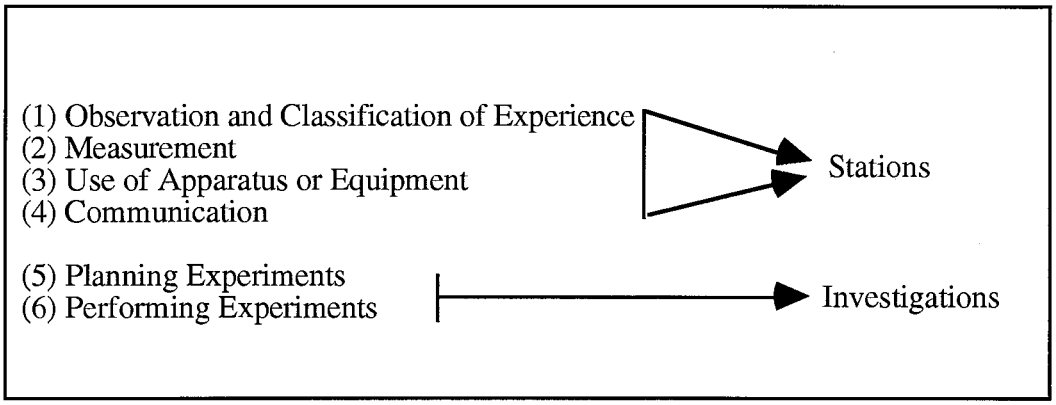


Figure 13. Mapping of Performance Assessment Dimensions to Task Type

In selecting stations and investigations for use in the assessment, the Contract Team applied several criteria. Those relevant to this focus question are shown in Figure 14.

Engagement :
Is the task interesting and likely to engage the student?

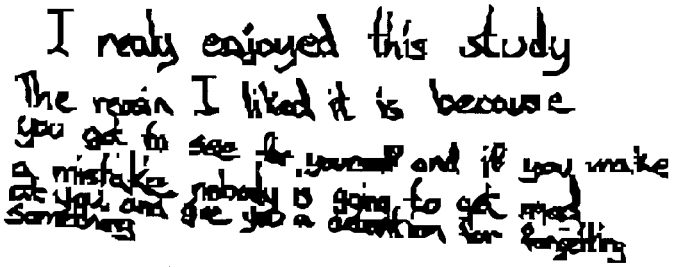
Appropriate to student abilities:
Does this task assess different abilities from the other stations?

Range of appropriateness:
Is the task suitable for more than one grade level?

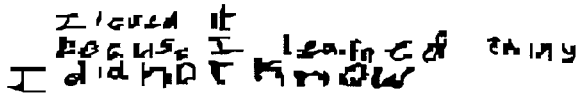
Figure 14. Criteria for the Selection of Tasks (Taken from Erickson et al., 1992, pp. 10 & 12)

Student Engagement

Student engagement as an explicit criterion for station selection might appear to be self-evident, but discussing the features of "good" tasks with students during pilot-testing reinforced the importance of this focus to the Contract Team. Comments collected from two Grade 4 students by a teacher during the assessment are shown in Figure 15.



"I realy enjoyed this study. The reason I liked it is because you got to see for yourself and if you make a mistake nobody is going to get mad at you and give you a detention for forgetting something"



"I loved it because I learned thing I did not know"

Figure 15. Grade 4 Student Comments about the Assessment Tasks

The positive comments of these two students were echoed by observations made by some of the teachers who had collected the data:

- ...we were not surprised by the comments from the kids: "You mean, this is science? I like **this** science." (female teacher);
- ...people were really positive and we did experience what you suggested might happen in one class where we didn't intend to test all of the kids -- we didn't need all of the kids. But the teacher said, "Oh, they'll feel really badly, please come back and do test them." (female teacher);
- ...we asked each student, would you rather do science this way or the conventional way? And the response was about 99% this way, so they were just thrilled with the idea. And, you know, we were thinking that we had never seen such enthusiasm for a math test or a reading test. (male teacher);
- They walked out of that room, nearly every one, thinking, "Oh, I'm a scientist". And I would love to try and capture that kind of similar experience in our classroom. It was a big thing I walked out of that situation with, this is great, and it was very special to have that time with those kids. (male teacher).

Very few students, typically less than 5%, did not respond to any part of a station task. Many of the stations elicited responses from 100% of the students, with several (Blocks and Boxes, Making Measurements and Paper Testing in Grade 4, Cool It, and Comparing Seeds in Grade 7; and Circuit Building and Cool It in Grade 10) leading to over 90% of students being judged as having performed at "satisfactory or better" levels.

A different strategy to engage students is manifested by the type of questions posed in the investigations. The instruction format on the student response sheet for the two investigations is similar; that for Magnets is shown as Figure 16.

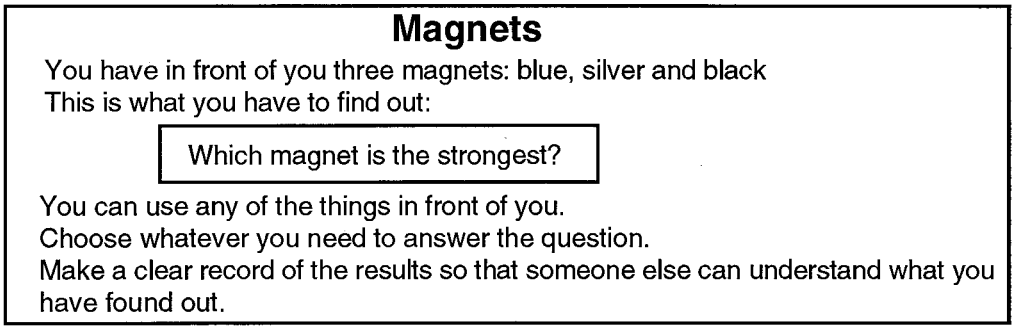


Figure 16. Student Instructions for Magnets Investigation

The question for Paper Towels is "Which kind of paper holds the most water?" (Gott and Murphy, 1987; Erickson et al., 1992). Such questions are very open-ended and allow students considerable latitude in their interpretation and experimental design. Analysis of the logical structures within the investigation tasks, and observation of students performing the investigations, enabled teachers and Contract Team members to identify the operational question that guided the students' experiments. For example, some students interpreted the paper towel question as "How much water can each paper physically hold (as a container)?" but others saw this question in terms of "How much water can be squeezed out of the wet paper?".

Assessing Different Abilities

In essence, this question addresses the face validity of the tasks. This is seen by Linn, Baker, and Dunbar (1991) as insufficient evidence upon which to hang the whole validation procedure. But, for the student with a different perspective towards the assessment process, the consequences of encountering tasks that are interesting and of educational value in their own right, must be considered. For example, the Grade 4 "Circuit A" saw students sorting beans, hitting strings on a sound board and listening to the sounds, testing simple materials for electrical conduction, moving a "puzzle box" to make inferences about the contents, listening to a personal cassette player to receive instructions for an experiment, and comparing how different pieces of wood float in water and salt water. These multi-sensory and tactile experiences were intended to convey the breadth of "hands-on" science to students and their teachers. Feedback from students and teachers indicate that this goal was achieved.

Range of appropriateness

The intent of this consideration was to demonstrate to teachers that growth in students' science abilities may be observed using a set of common tasks. Three stations (Sound Board, Rolling Bottles and Magnet Power), and two investigations, (Paper Towels and Magnets), were used at all three grade-levels. The stations were pilot-tested with students in Grades 4, 7 and 10. It was found that the Grade 4 students had greater problems in developing experiments to identify the stronger of the two magnets. Conversely many of the Grade 10 students completed the experiment quickly and accurately. In order to provide a differentiated degree of difficulty, the Contract Team decided to use pairs of magnets that were very different in strength for Grade 4 students, closer in strength for Grade 7 students, and very close for Grade 10 students; each magnet was within 5% of the mass of the other magnet and of the same dimensions.

Promotion of Models of Communication

The abilities to read and write are highly valued in our society and the written word remains the favoured medium for most formal education. As students progress through school there is the quite reasonable expectation that their abilities to communicate will improve, based upon developments in writing skills and an increase in vocabulary.

The data collection techniques for our stations depend completely upon the students being able to construct written responses to the questions. Coding and scoring of the stations was based solely upon the students' written responses. One station at Grade 4 (Blocks and Boxes), and a station used at both Grades 7 and 10 (Cool It), were devised to use a personal cassette player to present the instructions. A station common to all three grades (Magnet Power) requires students to draw a picture or diagram to show what they did in an experiment.

In the investigations there are three types of data: observation schedules completed by teachers, student descriptions, and teachers' notes of interviews with students at the conclusion of their investigation. Reliance upon student written responses is significantly reduced, as the principal source of descriptive data comes from the teachers through the observation schedules, and to a lesser extent through the interviews. This dependence on the teachers' descriptions of the students' experimental procedures is justified by comments from teachers at a debriefing session after the data collection:

> After the investigations and you're asking them questions, I always came back to the teacher in me, rather than the assessor and asking, "Have you ever done anything like this, this year?" And in every case the answer was "No". And then I asked, "If you had to write this up as a formal write-up, could you do that?" and the answer in every case was "No." This was Grade 7. So, they hadn't done anything like what we were doing with the magnets or the paper towels, which they really liked. (male teacher)

The teachers of Grade 10 students also reported that their students were unfamiliar with such open-ended experiments, but these students appeared better able to describe their procedures. The Grade 4 teachers chose to analyze the interview data as an oral report of the students' experimental procedures, an approach believed to be fairly consistent with classroom practice.

A further consideration of communication comes in the examination of the criteria used to describe each level of performance. These criteria were constructed for appropriate grade-levels by the teachers as they were coding the student responses. A consistent theme that emerged across all three grades was that extended responses were more highly rated. In the discussion of the results of the Grade 4 task "Sorting Beans" the Contract Team wrote:

> The teachers' judgments as to how well the students observed differences among the beans is puzzling. Sixty-one percent of students were judged to have performed at a satisfactory or better level, while only 1% were judged to have performed "extremely well." Teachers judged student performance by the number and complexity of the criteria they used. Category 3 required 1 criterion, category 4 required 2 criteria while Category 5 required 3 or more criteria. We think that "more is better" is a teacher judgement which disadvantages students. Some students may well value one salient criterion as an appropriate way to observe and designate similarities. It is in this adult perspective of appropriateness that the student performance may be undervalued. (Erickson et al., 1992, p. 19)

The Contract Team made similar observations at Grade 10 for the station "Sound Board":

> The Contract Team thinks that the criteria developed for this latter judgement have been too heavily weighted toward producing extensive explanations of the various relationships involved in this station which tends to undervalue those students who opted for more parsimonious explanations such as simply referring to differences in the vibrations of the strings. (Erickson et al., 1992, p. 143)

The message that students should engage in extended communication in responding to some of these tasks was clearly unanticipated by the Contract Team and perhaps reflects some differences in values. While such requirements may represent the practice in many classrooms, the Contract Team felt it appropriate to comment upon the apparent unfairness of some of the sets of criteria.

Content Analysis

Linn, Baker, and Dunbar (1991) consider that there are two components to be emphasized in considering the content of an assessment: first the coverage of the domain of interest by the content in the assessment, and second the quality of that coverage. These concerns are addressed by the question:

Are the assessment tasks appropriate for the students and within the curriculum?

British Columbia has a provincial curriculum for science. However, many educational initiatives, in particular those leading to the proposals outlined in *Year 2000: A Framework for Learning* (Ministry of Education, 1989) have led to three disparate curriculum documents: *The Primary Program* (Ministry of Education, 1990), *Elementary Science Curriculum Guide: Grades 1-7* (Ministry of Education, 1981) and *Junior Secondary Science Curriculum Guide* (Ministry of Education, 1985). The elementary science program in British Columbia therefore has two overlapping curriculum guides. The *Elementary* guide focuses upon curriculum materials that were already 15–20 years old in 1991, and are available in varying amounts in schools in British Columbia. Conversely, implementation of the relatively new Primary Program document suffered from a paucity of resources in 1991. Consequently, teachers of students up to Grade 4 were faced with the problem of presenting either the "old" or "new" curriculum with limited resources. Similar problems with the precision of the curriculum documents faced the teachers of Grade 10 students. The Contract Team was aware of these problems when choosing station tasks for the assessment. There was extensive cross-referencing to the curriculum documents to ensure that stations generally fitted with at least some provincial curriculum. Many more station tasks were presented to students for pilot-testing than were actually used in the assessment. Discussions with students, their teachers and teacher-members of an assessment review group led to the choice of stations that were used in the assessment. The traditional approach towards content validation was extended to include the voices and perspectives of students.

One of the tasks, "Rolling Bottles", produced very mixed reactions from teachers. Many who saw it were intrigued by the phenomenon of the bottle one-third full of sand rolling so slowly down the ramp. (When the bottles are rolled down the ramp, the bottle

that contain most sand rolls faster than less-full bottles; the one-third full bottle rolls slowest. When the empty bottle is introduced and rolls faster than partially filled bottles, the explanations become interesting.) The Contract Team presented this task to students in Grades 4, 7 and 10 with the intent of probing what students consider important in this situation. The Contract Team identifies the following 'objectives' for this station:

> 1. Students observe three bottles filled with different amounts of sand roll down the ramp, and identify the fastest.
> 2. Students explain why one bottle went fastest.
> 3. Students observe an empty bottle roll down the ramp and compare its speed with one bottle previously rolled.
> 4. Students explain why one of the two bottles went faster.
>
> (Erickson et al., 1992, p. 61)

This is a visually stimulating task and has been demonstrated by members of the Contract Team in seminars and at conference presentations. On two separate occasions, members of the audience, both with an interest in physics, have informed the presenter that this task is totally inappropriate for school students, perhaps even for university under-graduates! Both forcefully made the point that the physics of the task is so complex that no student would be capable of understanding the dynamics of the bottles. The presenters responded by pointing out that the intent of the task is to examine the sense that students make from their observations in this context. This type of debate about the quality of tasks demonstrates the significance of different values in evaluating both task quality and curriculum appropriateness.

Content quality and cognitive complexity must also be evaluated for the investigation tasks. The advantage of open-ended questions was discussed earlier in this chapter in terms of how the students were able to interpret questions at a range of levels. Analysis of the aggregated data of the students' work shows another effect:

> In both investigation tasks one is immediately struck with the apparent lack of differences between the students at the three age levels especially as it pertains to the planning and performance aspects of the tasks.

The Contract Team concludes that:

> ...these data demonstrate the capabilities of Grade 4 students to plan, to perform, and to interpret the results of open-ended investigations of this nature. Furthermore, it would seem that their experimental strategies are very similar to those adopted by the older pupils. Our findings indicate that the most common experimental approaches, at least for magnets and the absorption qualities of paper, are constructed by pupils at a much earlier age than many theorists or curriculum developers have predicted. Thus one curricular approach, which is firmly rooted in many science programs, is that younger pupils (say up to the age of 11 or 12) should be engaged in less complex activities than a complete investigation. The implicit, and some explicit, message is that younger students need to develop first the more basic "process skills" of science such as observing, classifying, measuring, and inferring before they can proceed to the more complex and sophisticated reasoning characteristic of conducting complete and valid investigations. Our data contradict this position. (Erickson et al., 1992, p. 243)

This interpretation of the data has important implications for curriculum planning, particularly in elementary science. It is consistent with perspectives articulated by Millar about general cognitive processes (1991), and Woolnough about the value of introducing elementary students to investigations (1989). Differences between younger and older students reported by the Contract Team are:

> ... the construction of more elaborate and powerful explanatory models that are used to frame experiments of the type that the teachers observed in this project...
> an increase in the abilities of students to perform adequately some experimental abilities such as conducting appropriate measurements with care and precision, identifying and controlling variables thought to be important to the outcome of the experiment, and finally providing an interpretation of the data and communicating that interpretation to others. (Erickson et al., 1992, p. 243)

Alternative interpretations of certain aspects of these data might lead to the following questions:

1. Did the investigations require sufficient curriculum content knowledge to enable the older students to show the depth of their knowledge?
2. Was the language of the questions appropriate to elicit cognitively complex responses from students capable of such responses?

The two investigation tasks were very similar in structure and provided convergent evidence which led to the inferences made by the Contract Team. However, it may be

argued that more complex tasks would have provided data that could more clearly identify the limits of performance of Grade 4 students and pose more of a challenge to the older students.

Instrumental Stability

When equipment is used in assessments it is necessary to ensure that the equipment used by one set of students will behave the same way over the time period of the assessment, and will also behave in the same way as that used by other sets of students. This is a vital consideration in the discussion of the reliability of the data collection procedures. The focus question for this part is:

Does the equipment behave consistently over time?

In preparing this assessment the Contract Team constructed 10 kits of equipment for each grade, including over 20,000 individual pieces of equipment. The Contract Team prepared the *1991 British Columbia Science Assessment Performance Tasks Administration Manual* (Bartley, 1991) which presents specific details of the equipment and procedures for set-up and administration. A small number of measurements of temperature (room and cold water) were recorded on "Administrator's Record Sheets" that were completed each time the station tasks were used. These sheets included the instruction "If anything happens, such as equipment malfunction or breakage, that you feel may affect the students' performance in this part of the assessment, please record on this sheet". Although over 50 rotations across the three grades were recorded on these sheets there were no reported cases of significant equipment malfunction. Typical problems included discharged batteries for the personal cassette players, but as spares were supplied this particular problem was minimized.

More recently the author has been involved in presenting the assessment tasks to teachers in over a dozen school districts around B.C. and at the 1993 National Science Teachers National Convention in Kansas City, Missouri. All of the equipment performed consistently with appropriate maintenance and replacement, e.g., batteries, calcium metal, pH paper, sandstone, vinegar, etc.

Administration Stability

There are two aspects of this perspective towards data collection. The first is related to the work of the Contract Team in providing clear directions for procedures to collect the data. The second is linked to the ability of the teachers, and the schools where the data were collected, to create similar environments in which students could work through the tasks. The focus question for this part of the inquiry is:

Are the administration procedures clearly developed and applied consistently?

The *1991 British Columbia Science Assessment Performance Tasks Administration Manual* (Bartley, 1991) was developed during the pilot-testing. In addition to the details of equipment discussed earlier in this paper, the *Manual* presents a set of procedures for administration of the assessment. These procedures include scripted orientation guides for teachers to read to the students while the assessment is under way. These scripts were used by the Contract Team and teachers during the preparation workshop. At the post-assessment debriefing, teachers declared that there were no difficulties in the application of these procedures during the assessment.

The data were collected at a time when the union that represented teachers in many school districts was in dispute with the Ministry of Education. Several of the teachers who administered the assessment reported problems in collecting data, and delayed going to

schools until the climate was more harmonious. A different problem arose where the school administrators were not able to find sufficient space. One teacher recounted his experience:

> we asked for two quiet places to work. The other teacher was OK. He got a whole room, big tables and everything. I got a little room, they call it the smoker's pit and I got two little tables about this wide to put my stuff on, a couple of little chairs.

But this response was countered by another who spoke of the assessment being:

> generally well-received. One school, they moved the kids that weren't writing our test to the library, so they opened up the classroom as well. We used it. They were very accommodating.

Discussions with the teachers suggest that few students were prevented from performing in a good working environment. Administrative stability appears to pose no threat to the reliability of the data collection.

Internal Consistency and Generalizability

This consideration has been a fundamental criterion in the evaluation of the reliability and validity of traditional forms of testing. Linn, Baker, and Dunbar (1991) consider that this continues to be important issue in the validation of alternative assessments. As the B.C. assessment was in the curriculum area of science, the factors that must enter the discussion originate in both the measurement and the science education communities. The focus question here is deliberately general, and interpretations will be articulated in the discussion:

What factors affect the generalizability of student performance across tasks in a science assessment?

There appear to be three distinct levels for this discussion. The most straightforward of these concerns the stability of the data collection procedures and has been discussed earlier

in this chapter. Next there is the issue of consistency in the coding and scoring procedures. Finally there is scope for conceptualization in how to make sense of the data.

Consistency in Coding and Scoring Procedures

Coding consistency was given high priority by the Contract Team. The requirement for consistency was emphasized to the teachers during the coding workshop, and procedures were instituted to examine the consistency of teachers' coding on all stations. Details of the procedures are presented the *Technical Report* (Erickson et al., 1992, pp.246-248), but a brief summary is presented here. Essentially, each team of four teachers coded the same five student response sheets; these sheets were then examined for matching pairs. Results are reported in terms of percentage of matching pairs, i.e., degree of consistency. A coefficient of 1 indicates 100% agreement of the teachers[1]. The coefficients for percentages of teachers in agreement with each other is reported for all codings on each station, as are results on codings for the judgements alone. Further analysis of the judgement data to identify scores that were only one point from agreement shows a generally high degree of consistency. Table 2 illustrates the inter-coder consistency for the six stations[2] in Circuit A from Grade 7.

---

1 If 3 out of the 4 teachers score the student at the same level, then 3 out of a possible 6 pairs are in agreement; the coefficient for inter-coder consistency is reported as 0.5.

2 These stations are chosen as a representative sample of the inter-coder consistency data generated by this process. The Grade 7 judgements are presented here as they were subjected to extensive analysis in the *Technical Report* (Erickson et al., 1992).

Table 2
Inter-coder Consistency Coefficients for Grade 7 Circuit A

Station	All codings for station	Judgements exact	Judgements ± 1 level
7.1 Guess It	0.97	0.88	0.12
7.2 Sound Board	0.73	0.48	0.32
7.3 Electrical Testing	0.89	0.65	0.23
7.4 Environmental Testing	0.88	0.80	0.20
7.5 Cool It	0.92	0.93	0.07
7.6 Floating Wood	0.72	0.62	0.30

These levels of inter-rater agreement are consistent with published acceptable levels for other performance assessments in science (Ruiz-Primo, Baxter, & Shavelson, 1993) and writing (Dabney, 1993). More recently Shavelson, Gao, and Baxter report that if raters are given appropriate training:

> The findings are consistent. Interrater reliability is not a problem. Raters can be trained to score performance reliably in real time or from surrogates such as notebooks [e.g., Baxter et al., 1992; Shavelson, et al., 1993].
> (Shavelson et al., 1994, p. 2)

Consistency Across Tasks

This is perhaps the most contentious issue in analysis of performance assessments in science. In most applications to this point, the number of tasks is usually too small for conventional reliability studies. Generalizability theory (Cronbach et al., 1972; Brennan, 1983) enables the user to "parcel out" variance to controllable facets which may contribute to the error (e.g., occasion, scorer, task). Variance arising from tasks is the area of interest in the examination of consistency of performance. Linn, Baker, and Dunbar identify generalizability theory as providing "a natural framework for investigating the degree to which performance assessment results can be generalized" (1991, pp. 18-19).

Extensive use of G theory has been made by the APU in England (Johnson, 1989), and in the U.S.A. by Shavelson and his associates (Baxter et al., 1992; Shavelson et al., 1992; Shavelson and Baxter, 1992; Shavelson et al., 1993; Shavelson et al., 1994). G theory has been characterized as "a statistical theory about the dependability of behavioral measurements". Within this definition, dependability is seen as

> the accuracy of generalizing from a person's observed test score on a test or other measure (e.g., behavior observation, opinion survey) to the average score that the person would have achieved **under all possible conditions that the test user would be equally willing to accept.** (Bold added; Shavelson and Webb, 1991, p. 1)

The application of generalizability theory requires that the test user accept the extension of conditions from the domain of the test to whatever domain or universe s/he wishes to generalize. Generalizations, particularly in science assessments, depend heavily upon construct labels and the definition of the domain of interest.

Rather than pursue a G study with the data, the Contract Team for the Student Performance Component of the 1991 B.C. Science Assessment chose to examine correlation matrices to see if students performed at consistent levels over different stations. The judgement questions for these stations are shown in Figure 17.

Station 7.1 Guess It
A. How well did the student estimate properties of length, mass, volume, area and time?
Station 7.2 Sound Board
A. How well did the student observe differences between the strings (visual and aural)?
B. How well did the student explain why the strings sound differently?
Station 7.3 Electrical Testing
A. How well did the student make inferences about the characteristics of conductors and non-conductors?
B. How well did the student predict/justify whether an object will be a conductor or a non-conductor?
Station 7.4 Environmental Testing
A. How well did the student draw inferences from collected data?
Station 7.5 Cool It
A. How well did the student listen and follow instructions orally?
B. How well did the student interpolate and extrapolate from data?
Station 7.6 Floating Wood
A. How well did the student observe how three wooden blocks float in water?
B. How well did the student predict what would happen to the third block in salt water (based on prediction and explanation)?

Figure 17. Judgement Questions for Grade 7, Circuit A (Erickson et al., 1992, p. 249)

The correlation matrix for Grade 7, Circuit A is shown as Table 3. The matrix is discussed in the *Technical Report*:

> In circuit A only 6 out of the possible 45 correlations are statistically significant (1-tailed test of significance, $p < .01$). Of these 6 correlations there are 4 that represent correlations between judgements within stations, i.e. from stations where there were two teacher judgements about different aspects of a student's performance on the same station. The strong correlations ($r > .5$ with $p < .001$) for this circuit are both within-station correlations, on Sound Board and Electrical Testing. (Erickson et al., 1992, p. 251)

Table 3
Pearson Correlation Matrix for Grade 7, Circuit A Data, N=115 students[3] (Bold added; Erickson et al., 1992, p. 249)

	7.1 A	7.2 A	7.2 B	7.3A	7.3 B	7.4 A	7.5 A	7.5 B	7.6 A	7.6 B
7.1 A	1.0	.004 p=.481	-.019 p=.421	.176 p=.030	.048 p=.307	.108 p=.125	.151 p=.053	.158 p=.046	.007 p=.469	.096 p=.156
7.2 A		1.0	**.648 p=.000**	.216 p=.011	.122 p=.099	.027 p=.389	.190 p=.021	-.117 p=.108	-.070 p=.229	.140 p=.068
7.2 B			1.0	.233 p=.006	.169 p=.036	-.053 p=.289	.141 p=.067	.012 p=.447	-.092 p=.163	.129 p=.086
7.3 A				1.0	**.544 p=.000**	.216 p=.011	.259 p=.003	.146 p=.060	.073 p=.220	.021 p=.414
7.3 B					1.0	-.047 p=.331	.102 p=.140	.106 p=.131	.027 p=.389	.039 p=.340
7.4 A						1.0	.210 p=.012	.169 p=.036	.038 p=.343	.150 p=.055
7.5 A							1.0	.226 p=.008	.098 p=.149	.096 p=.155
7.5 B								1.0	.101 p=.141	.018 p=.425
7.6 A									1.0	.234 p=.006
7.6 B										1.0

Uniformly, the within-station correlations are statistically significant; student performance was consistent within any given station. The Contract Team suggests that student performance in these stations is probably most influenced by "the task context (i.e., the type of content knowledge and experience embedded within the task)" (Erickson et al.,

---

3 Within-station correlations are shown within the confines of the "bold box".

1992, p. 251). Interestingly, correlations between student scores for judgements that purport to measure some aspect of a generalizable skill are very low:

> The most frequently discussed "skill" of science is "observation". In Sound Board students observe differences between strings, while in Floating Wood they observe how the wooden blocks float; the correlation is – .070. (Erickson et al., 1992, p. 251)

These results are consistent with the constructivist perspective that the Contract Team used in preparing the assessment framework. Millar and Driver present an eloquent argument against simply "process-based science" in their paper *Beyond Processes* (1987). The Contract Team supports this perspective in the *Technical Report:*

> We, at all costs, wanted to avoid the return in science teaching to the days of the Science as Process Approach where individual so-called "process skills" were taught separately. We believe that an emphasis on separate abilities or skills would encourage such an orientation. (Erickson et al., 1992, p. 253)

The evidence from the Student Performance Component of the 1991 B.C. Science Assessment and the work of Shavelson, Gao, and Baxter (1994) leads me to believe that the question of consistency across tasks depends upon the values (in the sense of a perspective about science) that drive the development of assessment tasks. If tasks are generally homogeneous **and** drawn from the same content domain, there is a strong likelihood of consistent student performance from one task to the next, i.e., the task scores will show a high degree of correlation and performance can be generalized, <u>but only to further similar tasks</u> (as for Shavelson et al., 1994). If the tasks are designed to be heterogeneous, covering a range of dimensions or domains (as in B.C.), then consistency of student performance should not necessarily be expected. Sets of heterogeneous tasks are more likely to be valid in representing a complex universe, say a science curriculum, than are sets of limited but reliable homogeneous tasks.

Fairness

The issue of fairness should be considered in any assessment, but is mandatory in an assessment that is intended to be sensitive to gender issues. The question focusing on fairness is:

Do the assessment data indicate any bias towards or against any specific, identified group?

The Contract Team attempted to be particularly sensitive to gender-related differences in student performance. There is consistent evidence to show that in the majority of science classes, fewer girls than boys handle science equipment (Kahle, 1988; Wienekamp, Jansen, Fickenferichs, & Peper, 1987). It was a concern of the Contract Team members that we present females and males with equal opportunities to demonstrate what they "CAN DO" by working through hands-on performance tasks. The pilot-testing indicated that the gender-related differences were subtle, and that only when complete data were collected and aggregated would differences, if any, appear. The stations were coded and scored using number identities; therefore the coders were not aware of the school location or the gender of the students[4]. In the assessment report, aggregated data are presented in tables like the one shown here as Table 4. The Contract Team chose to focus upon tasks for which differences in performance between males and females were greater than 15%. The reason for this choice is "pedagogical significance" (Erickson et al., 1992, p. 227) in that 15% represents a difference of three to five students in a class of 20 to 30 students. With this condition it is reported that "similarities in performance between males and females were more evident than the differences" (Erickson et al., 1992, p. 223).

---

4 Investigations were scored by analysis of the observation data, so the teacher/administrators were aware of the gender and identity of the students.

Table 4
Grade 10 Student Performance – Satisfactory or Better Rating by Gender, Stations 1 to 6 (Erickson et al., 1992, p. 226)

		Satisfactory or Better Rating		
		Total % N = 115	Female % N = 55	Male % N = 60
10.1	Guess It: How well did the student estimate properties of length, mass, volume, area and time?	30	23	35
10.2	Sound Board: A. How well did the student observe differences between the strings (visual and aural)?	77	77	77
	B. How well did the student explain why the strings sound differently?	41	36	44
10.3	Circuit Building: A. How well did the student build series and parallel circuits?	94	93	95
	B. How well did the student draw circuit diagrams?	94	95	94
10.4	Environmental Testing: How well did the student draw inferences from collected data?	65	**75**	**58**
10.5	Cool It: A. How well did the student follow instructions given orally?	98	100	96
	B. How well did the student draw a graph from data obtained from a table?	62	68	56
10.6	Microscope: A. How well did the student manipulate the microscope to obtain images at the stated magnifications?	34	29	38
	B. How well did the student calculate the increase in magnification?	10	10	10

Some differences are reported and discussed in the *Technical Report,* for example:

> Station 4, "Environmental Testing," produced a small gender difference at Grade 7, but at Grade 10 more females (75%) than males (58%) performed at a "satisfactory or better" level. The Classical Component of the British Columbia Assessment of Science Provincial Report (Bateson et al., 1992) has demonstrated that females show more interest and concern for the environment. We believe that this finding may point to a possible explanation for the observed differences in performance in this station. The abilities of males and females to measure and interpret pH values were similar. The quality of the females' explanations for the change in acidity of a lake was rated higher by the teachers. (Erickson et al., 1992, p. 227)

All three of the stations used across the three grades show interesting gender-related differences in performance. Performance on all of the stations shown in Table 5 favoured females at Grade 4. However, by Grade 10 both "Rolling Bottles" and "Magnet Power"

were seen to favour males. The Contract Team postulates that the differences for "Magnet Power" arose because of males' "prior interests and experiences in physics". However, if this is the case, why does this not manifest at an earlier grade? A focus upon the males' current interests and experiences in Grade 10, might be more convincing.

Table 5
Student Performance on Common Tasks by Gender
(Derived from Erickson et al., 1992)

Grade	Station	% Satisfactory or Better Rating					
		Grade 4		Grade 7		Grade 10	
		Fem. N=58	Male N=50	Fem. N=59	Male N=52	Fem. N=57	Male N=50
4	Making Measurements	99%	82%				
7 & 10	Instruments			34%	58%	55%	58%
4, 7 & 10	Rolling Bottles	71%	54%	28%	38%	44%	55%
4, 7 & 10	Magnet Power – Judgement A	67%	53%	76%	56%	64%	73%
4, 7 & 10	Magnet Power – Judgement B	52%	43%	76%	64%	76%	89%

As indicated earlier, the Contract Team chose to report only differences of 15% or greater. The number of judgements where the differences between females and males with satisfactory or better performance were greater than 15 % is shown in Table 6.

Table 6
Stations Where the Percentage of Females and Males Judged to have Satisfactory or Better Levels of Performance differs by more than 15% (Derived from Erickson et al., 1992, pp. 224-226)

Grade 4	Grade 7	Grade 10
2 judgements favouring females	1 judgement favouring females	2 judgements favouring females
0 judgements favouring males	1 judgement favouring males	1 judgement favouring males
Total judgements = 19	Total judgements = 19	Total judgements = 19

These data show that the number of judgements for which there is a large gender-related difference in performance is only one or two per grade out of a total of 19 judgements per grade. However, if differences of 10% are considered pedagogically significant the overall picture changes. These data are shown in Table 7.

Table 7
Stations Where the Percentage of Females and Males Judged to have Satisfactory or Better Levels of Performance differs by more than 10% (Derived from Erickson et al., 1992, pp. 224-226)

Grade 4	Grade 7	Grade 10
5 judgements favouring females	5 judgements favouring females	3 judgements favouring females
0 judgements favouring males	4 judgements favouring males	5 judgements favouring males
Total judgements = 19	Total judgements = 19	Total judgements = 19

At Grade 4 females appear to perform better than males; in Grades 7 and 10, both sexes appear to perform at similar levels. It is debatable as to whether the level of reporting should have been 10% rather than 15%. Perhaps what should be considered is the potential significance of the trend identified in Table 5. When relatively few students perform at satisfactory or better levels of performance then differences between females and males may be more important which would require consideration of smaller differences. Conversely, when the majority of students, female and male, perform a task at a satisfactory or better level of performance, it may be more reasonable to consider only larger differences between females and males.

Consequences

The most important component of this validation inquiry is addressed in drawing together earlier sections of this chapter to identify and examine the consequences of the assessment — intended or not. There are three concluding questions:

Were the intended consequences of the assessment achieved?

What were the unintended consequences?

What actions have been taken to support the "good" unintended consequences and abate the "bad" unintended consequences?

Intended Consequences Achieved

Explicit, intended consequences of the assessment were set out by the Ministry of Education and operationalized by the Contract Team. Discussion of the intended consequences of the assessment is presented here first in terms of the Ministry of Education and then from the standpoint of the Contract Team.

In 1992, the Ministry of Education published a document entitled *Ministry Response to the 1991 British Columbia Assessment of Science* (1992a) in which Ministry staff examine the findings and implications of the recommendations presented in the *Technical Report* (Erickson et al., 1992). The comments in this response document are directed towards the Contract Team's discussion of student performance in the investigation tasks, particularly the observation that students had experienced few opportunities in their classrooms to investigate in open-ended or self-directed ways. The Ministry responds by stating that it will encourage:

> ...science teaching that emphasizes greater student involvement with scientific inquiry and open ended scientific and mathematical problem solving. (Ministry of Education, 1992a, p. 8)

A subsequent draft document *Curriculum and Assessment Framework: Science* (Ministry of Education, 1992b) includes such a focus.

A further consequence of the assessment for the Ministry of Education has been the development of a school district-level package *Science Program Assessment through Performance Assessment Tasks: A Guide for School Districts* (Bartley et al., 1993). This self-contained binder contains details of all tasks, administration, data analysis and interpretation procedures enabling school districts to use the materials independently to collect their own district-level data.

The original, explicit agenda of the Contract Team is outlined in *The Assessment of Students' Practical Work in Science* (Erickson, 1990). This document indicates some of

the intended consequences that were goals for the Contract Team. The most significant of these is the viewing of student performance in science from a constructivist perspective on learning. The data provide empirical support for the Contract Team's position on two significant issues in science education: advocating an approach that might be described as "Investigations for All", and evidence that "processes" are context-dependent. One intended outcome was that students and teachers would support this mode of assessment; the actual degree of student and teacher enthusiasm surpassed our hopes.

Unintended Consequences

At this stage of the validation inquiry, several unintended consequences of the Student Performance Component of the 1991 B.C. Science Assessment have been perceived.

For the Ministry of Education one unintended consequence has been unwelcome administrative expenses. As the work of the assessment progressed, the labour-intensiveness of the whole enterprise became apparent. Data collection required trained administrators, as did the coding and evaluation of the data. The 36 teachers who were involved in the assessment each required 10 or 11 days of paid release time, and two visits to Vancouver for workshops. The cost of this teacher time and transportation can be considered as an undesirable consequence if one sets a low priority on teacher in-service education. However, many of the teachers reported that their work on this project was the most fulfilling in-service work of their careers – clearly a positive consequence!

A desirable although unanticipated consequence is the positive international recognition that this component of the Science Assessment has received (as part of the greater set of B.C. Science Assessment reports) including the honour of the AERA Division H Award for the Best Program Evaluation published in 1992. This award

recognized and rewarded the decision of the Ministry of Education to examine a broad range of variables in the 1991 B.C. Science Assessment.

From the perspective of the Contract Team, the unintended consequences identified to date have generally been seen as problems to solve rather than as unpleasant surprises. Details are presented in chronological order.

The first problem was related to task development and the procurement of resources. Initially it was expected that three investigations, one with a biological theme, would be used in the assessment. Investigations that have been used by the Assessment of Performance Unit (APU) in England (Gott and Murphy, 1987) entail the use of African land snails, and mealworms or woodlice (sowbugs). Inquiries about sources of African land snails soon led to the discovery that this animal is banned from Canada because of a perceived agricultural risk. As this part of the pilot work took place in the late fall of 1990 and early spring of 1991, local garden snails were not available because of winter hibernation. Instead, snails were purchased from a grower of edible snails, and plans were made for pilot testing. Lengthy observations of these snails, a tropical variety, indicated that their movement was significantly slower than the local garden snails, but increased as the ambient temperature was raised. Only when their environment was above 25°C were these snails active enough to be useful in an assessment situation. As this presented an unworkable constraint, this investigation was not used. The Contract Team found similar problems with the activity of mealworms. Consequently, the decision was made to use only the two investigations "Magnets" and "Paper Towels" and forgo the biological theme.

The next problem came about when the Contract Team attempted to find a science equipment supplier willing to contract for the construction of the assessment kits. While the accessible suppliers were confident of their ability to provide the range of materials required, they were not willing or staffed to construct the kits. This work fell to the Contract Team and led to many extended workdays. It was, however, completed on time.

The teachers who administered the assessment appreciated the work and made comments such as:

> ...my feeling was, and I think the other teacher from my district agrees, is that the job you did setting it up for us was tremendous. Ninety-nine percent of the problems were solved before they even got underway. (male teacher, July 1991)

A third problem arose with some of the teacher-developed criteria for the evaluation of student performance. A consistent theme running through these criteria for a noticeable number of stations was that teachers valued extended or multiple responses from the students. This presented a problem for the Contract Team as the criteria had been developed by a set of teachers with extensive experience not only of teaching appropriate grade-levels in B.C., but also in this mode of assessment. In its discussion of specific stations in the *Technical Report*, the Contract Team comments upon its perception of the apparent unfairness of such criteria, and the consequences in the interpretation of the data.

The discussion of gender-related differences in the *Technical Report* concludes that while there may be some task-dependent differences in the performance of females and males, overall results show that "gender-related differences are noteworthy by their absence" (Erickson et al., 1992, p. 228). The choice of 15% as the critical difference to guide the analysis, rather than a careful examination of the context for each of the differences, influenced the interpretation of the data. The potential for revisions to this analysis of gender-related differences remains.

The benefits of extensive involvement of teachers from around the province were seen as an unintended outcome of the assessment program by the Contract Team. Teacher involvement had been planned as part of the data collection procedures originally proposed to the Ministry of Education, but such a positive and extended interaction was not anticipated.

Because the Ministry of Education concluded that the assessment tasks were valuable exemplars, the Contract Team received a further contract from the Ministry of Education. This involved members of the Contract Team in field testing and producing *Science Program Assessment through Performance Assessment Tasks: A Guide for School Districts* (Bartley et al., 1993). This opportunity for students and teachers across B.C. to work with the station tasks and gain further experience with assessment procedures must be considered another positive consequence. It does however, mean that this set of tasks cannot be used in the future to measure change in student performance!

Refinements in Assessment Procedures

Work on field testing *Science Program Assessment through Performance Assessment Tasks: A Guide for School Districts* (Bartley et al., 1993) enabled the Contract Team to reflect upon its experiences, and to address concerns that had arisen in the Student Performance Component of the 1991 B.C. Science Assessment. Most significant of these was the decision to review, and encourage rewriting of, some teacher-derived criteria for student performance. These revisions were done by practicing teachers, with more guidance from members of the Contract Team than had been given to the teachers who produced the original criteria.

The orientation workshop for using *Science Program Assessment through Performance Assessment Tasks: A Guide for School Districts* (Bartley et al., 1993) includes discussion of the explicit and implicit messages about science that teachers and/or students perceive while performing the tasks, and also encourages debate about the implications for student learning of using these tasks.

REFLECTIONS UPON VALIDATION QUESTIONS

The approaches to validation inquiry set out in this chapter represent a move forward in the identification and resolution of some of the issues in validation of performance assessments in science. The questions asked in this study are derived from Shepard's fundamental question "What does the testing practice claim to do?" (1993, p. 429), and were influenced by the perspectives presented by Linn, Baker, and Dunbar (1991). However, this study has focused not only upon the claims and recommendations made from analyses of test responses, but has included an examination of the theoretical perspectives used to develop and interpret the assessment. Examination of the theories underlying test interpretation is a critical issue for validation (Messick, 1989b, 1994). Inclusion of the structural aspects of test development represents an important broadening of the validation process.

The theoretical perspective towards learning for the Student Performance Component of the 1991 B.C. Science Assessment is identified as constructivist in *The Assessment of Students' Practical Work in Science* (Erickson, 1990). This perspective permeates the approach taken by the Contract Team throughout the assessment from the early stages of task development and pilot studies through to data analysis and recommendations to the Ministry of Education as the project was nearing completion. Inconsistencies arose when different perspectives were brought to bear upon the interpretation of the data. For example, as noted previously, certain teacher-developed criteria in the station tasks appear to value extended but simple responses over shorter answers that indicate depth of understanding. As a result of this, the Contract Team ensured that more consistent perspectives were applied to criterion development for use in the *Science Program Assessment through Performance Assessment Tasks: A Guide for School Districts* (Bartley et al., 1993).

"Validation is essentially a matter of making the most reasonable case to guide both current use of the test and current research to advance understanding of what the test scores mean" (Messick, 1989b, p. 13). Many of the factors which tend to reduce or obscure the validity of inferences have been addressed by the Contract Team. At least two more issues remain: 1) gender-related differences in student performance require further scrutiny, and 2) the design of cognitively complex tasks, stations and investigations, for all grade levels, should be discussed. Validity is "a matter of degree, not all or none" (Messick, 1989b, p. 13). This inquiry has identified some specific difficulties with the inferences and theoretical rationales made by the Contract Team. However, my conclusion from the evidence is that in general the assessment did achieve the intended goals.

CHAPTER 5 — DESCRIBING STUDENT PERFORMANCE IN SCIENCE

The issue of how to describe student performance in hands-on science tasks is discussed in this chapter which constitutes my response to the second research question, "What are the essential characteristics of descriptors of student performance on performance tasks in science?" The characterization of a task, or set of tasks, is that vital element in the set of inferences that would typically be examined during validation. Cherryholmes (1989) summarizes the initial status of construct validity as "the *identity* between an attribute or quality being measured and a theoretical construct" (Emphasis in original; p. 100). This attempt to provide a correct description of the world was soon found to be impractical. Subsequently, construct validity was redefined as "the entire body of evidence, together with what is asserted about the test in the context of the evidence" (Cronbach and Meehl, 1955, p. 284; cited in Cherryholmes, 1989, p. 101). This shift in definition moved construct validity from a defining identity to a defensible interpretation[1].

For assessment in school science the issue is multi-faceted at both the macro and micro levels. At the macro level there is a need to examine the identity of the construct "school science" as it pertains to current curricula. At the micro level the issue becomes one of task description. There must be congruence between the micro and the macro; the domain of items chosen for the assessment tasks must provide an acceptable or defensible representation of the universe of "school science".

The perspectives among science educators that have guided curriculum and assessment are reviewed. The chapter continues with an examination of the frameworks

---

1 The various sets of *Standards for Educational and Psychological Testing* describe what is accepted as evidence in the argument to defend the interpretation.

that have been developed to provide descriptions of hands-on performance assessment items in school science. A model for the description of performance tasks is then presented.

SCHOOL SCIENCE

The parallels between assessment design and curriculum design should not be understated as each of these is informed by a view of "What counts as science education?" (Roberts, 1988, p. 27). Roberts focuses upon the tensions between the shaping of science education by policy makers and that done by classroom teachers. He identifies three inherent aspects of the question "What counts as science education?":

> First, the answer to it requires that choices be made – choices among science topics and among curriculum emphases. Second, the answer is a defensible decision rather than a theoretically determined solution to a problem theoretically posed. Third the answer is not arrived at by research (alone), nor with universal applicability; it is arrived at by the process of deliberation, and the answer is uniquely tailored to different situations. Hence the answers to the question will be different for every educational jurisdiction, for every duly constituted deliberative group, and very likely for every science teacher. (Roberts, 1988, p. 30)

The negotiations as to what should constitute an appropriate policy for science education in any particular jurisdiction are complex and require long periods of discussion and refinement. For example, this has been ongoing for almost ten years in the English National Curriculum, and is approaching four years for the drafts of the National Curriculum Standards for Science in the United States. Roberts argues that individual teachers who receive the "authorized" view of science education in the form of a curriculum or syllabus will maintain their own opinions about what "really" counts as science education. Consequently, there will be significant variations in emphasis from classroom to classroom (Roberts, 1988).

The view of science presented by formal curriculum documents has served to direct the development of student assessments. The developers of the *British Columbia*

Assessment of Science 1991 Provincial Report (Bateson et al., 1992) demonstrate this by identifying "authorized" emphases for science education[2] in the Province, where:

> An examination of recent curricula and general education documents in British Columbia reveals that the principal intention of science education is to foster "scientific literacy" through:
>
> - developing a positive attitude toward science;
> - expressing scientific attitudes;
> - developing science skills and processes that allow for exploration and investigation of the natural world;
> - understanding and communicating principles and concepts that provide a scientific perspective on the world;
> - appreciating cultural and historical contributions to science; and
> - understanding how science, technology, and society interact and influence one another, creating socioscientific issues.
>
> (Bateson et al., 1992, p. xvi)

Multiple perspectives as to what should be emphasized in science education are represented in these curriculum documents. As these statements relate to science education for all students in the public school system in British Columbia, such a range in emphasis is expected, and appropriate.

In planning the 1991 British Columbia Assessment of Science, the Ministry of Education recognized that multiple-choice questions alone would not provide sufficient breadth of information from which to make judgements about science education in the Province. Therefore, as described in Chapter 3, separate contracts were developed for the

---

2 This intention has also been stated by the Ministry of Education as the four goals of the science program (Bateson et al., 1992, p. xvi):

Goal A: The Science Program should provide opportunities for students to develop positive attitudes toward science.

Goal B: The Science Program should provide opportunities for students to develop the skills and processes of science.

Goal C: The Science Program should increase students' scientific knowledge.

Goal D: The Science Program should provide opportunities for students to develop creative, critical and formal (abstract) thinking abilities.

four components of the assessment[3] . The student performance component was intended by the Ministry of Education to focus upon "science skills and processes that allow for exploration and investigation of the natural world" (Bateson et al., 1992, p. xvi). This aspect of the assessment appears to be consistent with the curriculum emphasis that Roberts categorizes as "scientific skill development" (Roberts, 1988, p. 45). Roberts surveyed science education practice in elementary and secondary schools in North America and identified seven emphases. These, together with the seven underlying views of science, are shown in Figure 18.

Curriculum Emphasis	View of Science
Everyday Coping	A view of meaning necessary for understanding and therefore controlling everyday objects and events.
Structure of Science	A conceptual system for explaining naturally occurring objects and events, which is cumulative and self-correcting.
Science, technology, decisions	An expression of the wish to control the environment and ourselves, intimately related to technology and increasingly related to very significant societal issues.
Scientific skill development	Consists of the outcome of correct usage of certain physical and conceptual processes.
Correct explanations	The best meaning system ever developed for getting at the truth about natural objects and events
Self as explainer	A conceptual system whose development is influenced by the ideas of the times, the conceptual principles used, and the personal intent to explain.
Solid foundation	A vast and complex meaning system which takes many years to master.

Figure 18. Seven Curriculum Emphases and Associated Views of Science (Roberts, 1988, p. 45)

---

3 These are:

Component 1: The Classical Component

Component 2: The Student Performance Component

Component 3: The Socioscientific Issues Component

Component 4: The Context for Science Component

The "skill development" perspective, with its dependency upon the processes of science, has guided the development of other performance assessment projects. Of these, the Second International Science Study (SISS) has been most prominent, with a report entitled *Assessing Science Laboratory Process Skills at the Elementary and Middle/Junior High Levels* (Kanis, Doran, & Jacobson, 1990).

In stating as an underlying view of science the "usage of certain physical and conceptual processes" (1988, p. 45) Roberts does not pursue the description of the set of processes which are included in such a perspective. Donnelly and Gott (1985) consider that a coherent structure for science processes is an attractive approach for science educators to use in assessment, as it appears to present a unified perspective of science through an assessment framework.

ASSESSMENT FRAMEWORKS

The mechanics of using a curriculum as the starting point for science assessments in British Columbia is illustrated in the *1986 British Columbia Assessment of Science General Report* (Bateson et al., 1986). The diagram used to represent the chronology of development is shown as Figure 19.

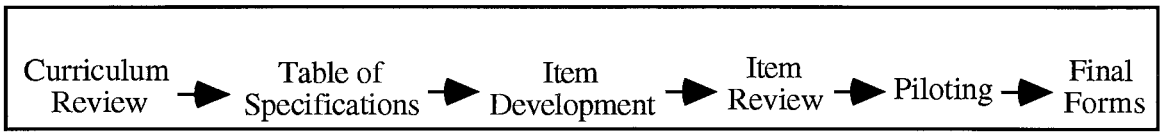


Figure 19. Development of the Achievement Instruments (Bateson et al., 1986, p. 8)

The steps shown in Figure 19 identify procedures which attempted to "ensure that the items used in the assessment were of the highest quality possible and that the overall survey accurately reflected the appropriate curriculum" (Bateson et al., 1986, p. 8). The second step, development of Tables of Specifications for each grade, plays an important part in translating the curriculum into a set of items used in the assessment by giving formal

definition to the relationship between curriculum content and assessment items for each grade level. In addition, the Tables of Specifications act as the reference point for the characterization of each item. The technique of examining items for "goodness-of-fit" within a Table of Specification, by panels of experts, has long been a part of the formal validation process. Bateson and his collaborators in the Classical Component of the 1991 British Columbia Assessment of Science (Bateson, Anderson, Brigden, Day, Deeter, Eberlé, Gurney, & McConnell, 1992) extended this methodology to include "think aloud" interviews with students about the individual assessment items.

For "hands-on" performance assessment, the Table of Specifications is usually replaced or augmented by an assessment framework. Assessment frameworks serve a variety of purposes — from describing the tasks or items at one extreme, to attempting to configure thinking processes at the other. The type of framework, its structure, and its use will vary significantly depending upon the purposes and values of the assessment developers. The frameworks discussed here represent a wide variety of viewpoints ranging from the pragmatic APU approach to the higher-order thinking skills framework of NAEP (Blumberg et al., 1986). My examination of the International Association for the Evaluation of Educational Achievement (IEA) projects looks at the Second International Science Study (SISS) and the Third International Mathematics and Science Study (TIMSS). I conclude with an examination of the framework used in the Student Performance Component of the 1991 B.C. Science Assessment (Erickson et al., 1992).

The APU Framework

Murphy (1990) reflects upon the framework used by the APU:

> It is problematic to select the terms by which pupils' attainment in science may be described. No single philosophical or psychological model was found to be an appropriate basis for the assessment exercise and its defined purposes, so none is reflected in the science activity categories. Nor was any hierarchy implied in the list. The activities are often referred to as synonymous with science processes. At other times the link is more cautiously stated. Although many generally accepted process terms (e.g., interpreting, observing, and hypothesising) are represented in either the category titles or in specific question descriptors in the categories, the science activities do not define the science processes. Moreover, the operational definitions of the activities include far more components of performance than are normally covered in discussions of the 'processes,' e.g., identifying the status of variables in investigations. (p. 153)

This cautious approach led the APU to develop a framework with six general category statements. These are:

1. Using symbolic representations
2. Use of apparatus and measuring instruments
3. Using observations
4. Interpretation and application
5. Design of investigations
6. Performing investigations (Johnson, 1989, p. 11)

Extensive lists of scientific concepts and knowledge were developed for use in the assessment of students at age 11 (Russell et al., 1988), at age 13 (Schofield et al., 1988) and at age 15 (Archenhold et al., 1988). The work of the APU took place in a relatively unstructured setting before the development of the National Curriculum in England. The freedom to select any focus allowed the APU management to adopt a view of science as "an experimental subject concerned fundamentally with problem-solving" (Murphy and Gott, 1984, p. 5). Murphy and Gott believe that this decision led to a framework defined by three dimensions:

- the science process involved in answering the question
- the degree of conceptual understanding required for its solution
- the content of the question and the context in which it is set

(1984, p. 5)

These dimensions were used by the APU both in describing the stations and also in the reporting of the results. Stations used by the APU were reviewed by external validators — practicing heads of science — with each station assigned to a unique sub-category as a uni-

dimensional task. Murphy reports that each station description includes "what the pupil was given, the expected outcome, and the mode of question response" (Murphy, 1990, p. 153).

Reporting of student performance using these three dimensions was quickly rejected by the APU management as too expensive and time consuming. The use of six framework categories, three content categories, and three context categories (science, other subjects and everyday) would lead to a grid of 54 cells! Johnson (1989) describes the effect of simplifying the reporting procedures as "reduced score interpretability". When reporting using the context dimension was abandoned, the remaining matrix consisted of 18 cells and this was considered still too complex. The problem was resolved by the decision that:

> the questions developed in any Category should be designed to be free of any dependence on taught science concepts. (Johnson, 1989, p. 11)

One consequence of this decision is that student performance is reported with no reference to science content!

The NAEP Project: A Pilot Study of Higher Order Thinking Skills Assessment Techniques in Science and Mathematics

An alternative model framework is presented by Blumberg, Epstein, MacDonald and Mullis in the NAEP report *A Pilot Study of Higher Order Thinking Skills Assessment Techniques in Science and Mathematics* (1986). Identified as a model for higher order thinking skills in science and mathematics, the framework is based on the premise that:

> at the most general level, higher order thinking skills are used to formulate a question, design and perform an analytical procedure and reach a conclusion to a problem. Further, such thinking was considered to be continuously self-monitored and evaluated as it occurs during the course of working through a problem or situation. Finally, subject-matter knowledge, beliefs, and values also impact upon how effectively an individual employs thinking skills in a particular situation. (Blumberg et al., 1986, p. 11)

The model, shown here as Figure 20, is built around six "aspects" that are described by the authors as representing what is "done" in science and mathematics. These aspects are intended to "collectively comprise the complex network of thinking skills in science and mathematics" (Blumberg et al., 1986, p. 11).

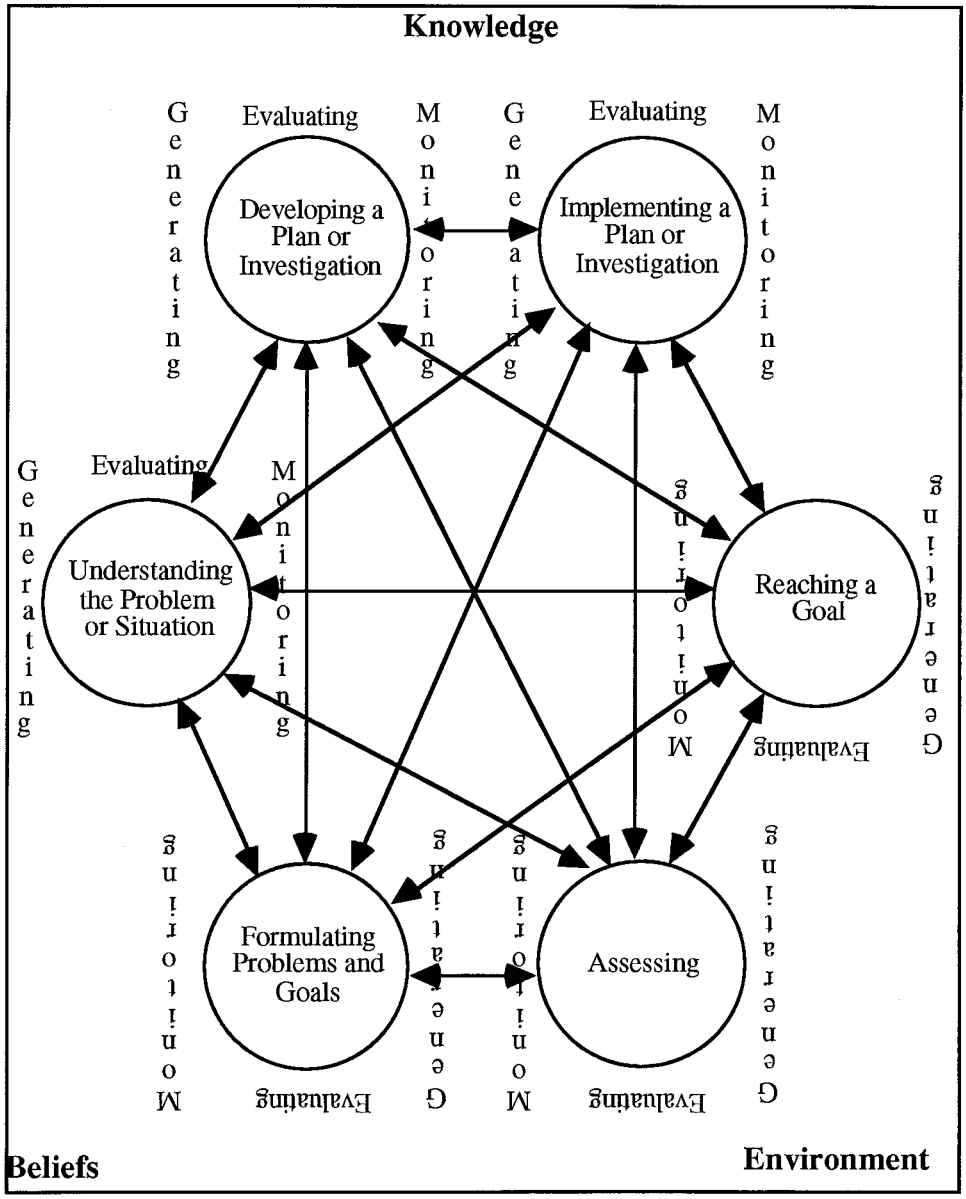


Figure 20. Higher Order Thinking in Science and Mathematics (Blumberg et al., 1986, p. 12)

The framework also invokes the "thinking skills" of generating and evaluating/monitoring, with the additional contextual demands of knowledge, beliefs and the environment[4].

Linking each "aspect" with each of the other five is intended to demonstrate the "dynamic" nature of the process. The authors recognize that the "distinctions between the aspects are fuzzy" (Blumberg et al., 1986, p. 15). This "fuzziness" is problematic when using this model, particularly as students reformulate problems to accommodate their own views of science, and their perceptions of the problem. Such models for higher order thinking skills, and similar models for problem solving, often provide "classical" representations which merely approximate the actual procedures. The influence of tacit knowledge is usually so powerful that students have difficulty in describing what they have done and why they have done it (Woolnough and Allsop, 1985). This framework for higher order thinking skills was conspicuous at the initial stages of the NAEP project, but received less attention after the APU personnel became involved. This framework was not even discussed in the concluding section of the report (Blumberg et al., 1986). As the purpose of the NAEP project was to evaluate only the feasibility of the procedures, the measurement properties of the assessment tasks were not published.

---

4 The environment for these considerations can be seen to include "external parameters such as recent experiences, working conditions, and testing situations, as well as internal parameters such as personality characteristics and interpersonal reactions. These environmental parameters may affect interest, motivation, attitudes, involvement, perseverance and cooperation" (Blumberg et al., 1986, p. 13).

IEA Projects

Two IEA projects are considered here -- the Second International Science Study (SISS) which ran from 1982 to 1992, and the Third International Mathematics and Science Study (TIMSS) which started in 1991 and will run beyond the end of this century.

The Second International Science Study (SISS)

When reporting the results and procedures of the hands-on component of SISS[5] in the United States, Kanis, Doran, and Jacobsen (1990) recognize the influence of the course *Science – A Process Approach* (AAAS, 1965) on science education in the United States. Kanis, Doran and Jacobsen report discussions of the IEA planning committee in 1983 in which several taxonomies of process skills were examined. Foremost among these was the Table of Specifications developed by Klopfer (1971) which was used for the SISS curriculum analysis. The APU framework (Murphy and Gott, 1984), and a list of skills developed by Tamir and Lunetta (1978) were also considered. The planning committee chose to adopt a three-category system, and justified this choice by stating that such a system "readily related to existing process skill schemes and was simple to use and explain" (Kanis et al., 1990, p. 14). The planning committee recognized that this framework did "not include knowledge and understanding as separate categories, since elements of cognitive outcomes are present in all items" (Kanis et al., 1990, p. 14). The SISS categories are shown in Figure 21.

---

5 For SISS, Population 1 consisted of students aged 10 (grade 5) and Population 2 was made up of students aged 14 (grade 9).

Skill	Illustrative Component
Performing (P)	To include: observing, measuring, manipulating
Investigating (I)	To include: planning and design of experiments
Reasoning (R)	To include: interpreting data, formulating generalizations, building and revising models

Figure 21. SISS Process Test Skill Categories (Kanis et al., 1990, p. 14)

Kanis, Doran, and Jacobsen relate that the coordinators from the six countries participating in SISS scrutinized the set of items chosen for the assessment and "assigned a skill classification to each item" (1990, p. 14). The issue of content representation was addressed by members of National Committees and science educators from participating countries using the SISS curricular grids. Tasks in SISS were characterized by both an identified skill and a content area such as Chemistry, Biology or Physics. In a post hoc analysis (Tamir, Doran, & Chye, 1992) the SISS performance tasks are mapped onto the Klopfer scheme, enabling comparison of the hands-on tasks and paper-and-pencil tasks. This post hoc analysis shows that these tasks were perceived as multi-dimensional; each task for 10-year-olds involved four to six process skills, while the tasks for 14-year-olds were assigned seven to nine process skills.

The complexity of describing of performance tasks in science is exemplified by the approach taken for SISS. Tamir, Doran, and Chye (1992) report the procedures that were followed in the development, administration and scoring of the SISS performance tasks. These are shown in Figure 22.

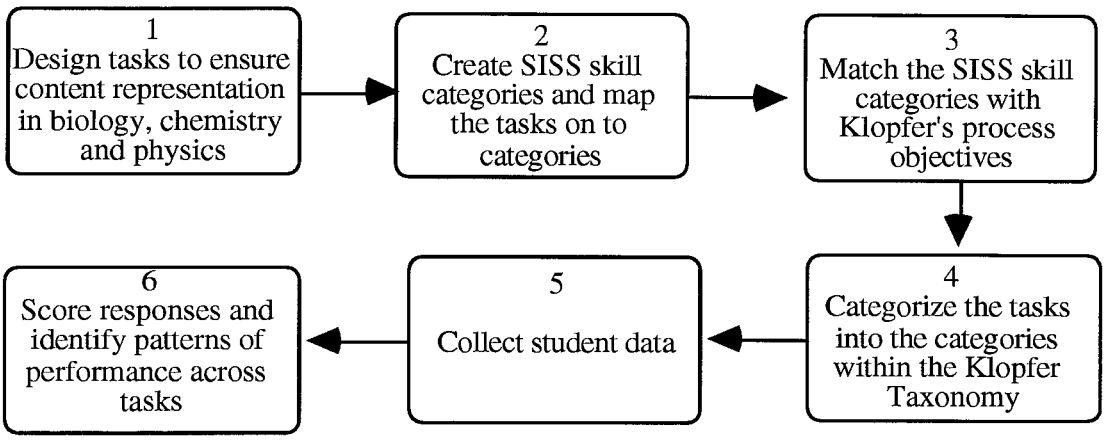


Figure 22. SISS Procedures

These six steps, from task development through to interpretation, allow for the description of task attributes. Some problems arise in the consistency of labelling, for example, where SISS practical skills categories and Klopfer's taxonomy overlap (Figure 23).

SISS Category	Klopfer's areas and skills	
Performance	B.1	Observation of objects and phenomena
	B.2	Description of observations using appropriate language
	B.3	Measurement of objects and changes
	G.1	Development of skills in using common laboratory equipment
Investigating	C.3	Selection of suitable tests of a hypothesis
	C.4	Design of appropriate procedures for performing experiments
Reasoning	D.1	Processing of experimental data
	D.3	Interpretation of experimental data and observations
	D.5	Evaluation of a hypothesis under test in the light of data obtained

Figure 23. Correspondence between SISS Practical Skill Categories and Klopfer's Scheme (Tamir et al., 1992)

The category "Performance" in the SISS scheme was seen to map onto four distinct skills in Klopfer's taxonomy (Figure 24, next page). However, the examination of individual tasks led to the identification of sets of skills beyond those that Tamir, Doran and Chye had identified as corresponding between framework and taxonomy.

A.0 Knowledge and Comprehension
- A.1 Knowledge of specific facts
- A.2 Knowledge of scientific terminology
- A.3 Knowledge of concepts of science
- A.4 Knowledge of conventions
- A.5 Knowledge of trends and sequences
- A.6 Knowledge of classifications, categories and criteria
- A.7 Knowledge of scientific techniques and procedures
- A.8 Knowledge of scientific principles and laws
- A.9 Knowledge of theories or major conceptual schemes
- A.10 Identification of knowledge in a new context
- A.11 Translation of knowledge from one symbolic form to another

B.0 Process of Scientific Inquiry I: Observing and Measuring
- B.1 Observation of objects and phenomena
- B.2 Description of observations using appropriate language
- B.3 Measurement of objects and changes
- B.4 Selection of appropriate measuring instruments
- B.5 estimation of measurement and recognition of limits in accuracy

C.0 Process of Scientific Inquiry II: Seeing a Problem and Seeking Ways to Solve it
- C.1 Recognition of a problem
- C.2 Formulation of a working hypothesis
- C.3 Selection of suitable tests of a hypothesis
- C.4 Design of appropriate procedures for performing experiments

D.0 Process of Scientific Inquiry III: Interpreting Data and Formulating Generalizations
- D.1 Processing of experimental data
- D.2 Presentation of data in the form of functional relationships
- D.3 Interpretation of experimental data and observations
- D.4 Extrapolation and interpolation
- D.5 Evaluation of a hypothesis under test in the light of data obtained
- D.6 Formulation of generalizations warranted by relationships found

E.0 Process of Scientific Inquiry III: Building, Testing and Revising a Theoretical Model
- E.1 Recognition of the need for a theoretical model
- E.2 Formulation of a theoretical model to accommodate knowledge
- E.3 Specification of relationships satisfied by a model
- E.4 Deduction of new hypotheses from a theoretical model
- E.5 Interpretation and evaluation of tests of a model
- E.6 Formulation of revised, refined and extended model

F.0 Application of Scientific Knowledge and Methods
- F.1 Application to new problems in the same field of science
- F.2 Application to new problems in a different field of science
- F.3 Application to problems outside of science (including technology)

G.0 Manual Skills
- G.1 Development of skills in using common laboratory equipment
- G.2 Performance of common laboratory techniques with care and safety

Figure 24. Klopfer Table of Specifications for Science Education (from Tamir et al., 1992)

For example, task 2A1 for Population 2, "Electrical Circuits" is a task in which students "determine the circuit within a black box by testing with battery-bulb apparatus" (Tamir, Doran, Kojima, & Bathory, 1992, p. 279). Figure 25 shows how this task was elaborated further, using both the SISS Content Classification Scheme (Postlethwaite and Wiley, 1991) and Klopfer's Taxonomy of Process Skills:

Task	Content	Klopfer's Process Skills
2A1	P53 Current Electricity	A1, A2, A3, A4, B1, D1, D6, G1, G2

Figure 25. Classification of Exercise 2A1 (Tamir, Doran, Kojima, & Bathory, 1992)

"Electrical Circuits" consists of three questions or items; so it is possible to identify the SISS skills for each part (Figure 26):

Task Part	SISS Skill	Content
2A1.1	Performing	Physics
2A1.2	Performing	Physics
2A1.3	Reasoning	Physics

Figure 26. Classification of Task 2A1 by SISS Practical Skill Categories (Tamir, Doran, Kojima, & Bathory, 1992, p. 281)

Figures 25 and 26 show the breadth and depth of the scrutiny to which the SISS tasks were subjected. Tamir and Doran confirm some of the difficulties of the categorization procedures employed by SISS.

> It was evident that placing an item within one skill area was necessary for analysis purposes, but this approach over-simplified the holistic, multi-faceted nature of many of the items. (Tamir and Doran, 1992b, p. 145)

Tamir and Doran report that "for both test forms, at both population levels, the success rate was highest for the performing skill and lowest for the reasoning skill" (1992b, p. 395). Task structure tended to have "reasoning items" following other skill areas. This led Tamir

and Doran to present an interpretation based upon a probable hierarchical structure. They propose that reasoning is the higher order skill, requiring successful use of lower order skills. Tamir and Doran (1992b) interpret these results as supporting similar claims for a hierarchy of science skills made by Bathory (1985).

In SISS, the use of a relatively simple three-state set of skills categories, combined with three content area descriptors, allows construction of a three-by-three matrix as shown in Figure 27.

	Content Area		
Skill	Biology	Chemistry	Physics
Performing			
Investigating			
Reasoning			

Figure 27. SISS Skill/Content Matrix

While there were some explorations in the use both of Klopfer's taxonomy and the SISS Content Classification Scheme (Tamir, Doran, Kojima, & Bathory, 1992), the SISS analysis (Tamir and Doran, 1992b) focuses upon performance as categorized within the matrix shown here as Figure 27. The means of the correlation coefficients between skills, for all students in both populations are positive, but low as shown in Table 8.

Table 8
SISS Skill Category Mean Correlation Coefficients

	Population 1	Population 2
Performing/reasoning	0.25	0.63
Performing/investigating	0.27	0.32
Investigating/reasoning	0.29	0.30

Doran and Tamir offer as a possible interpretation that "these subtests were not measuring the same skills" (1992, p. 371). These authors also warn that "the low correlations might

be due in part to the substantial errors of measurement in the assessment of the practical skills tests" (1992b, p. 371). One must consider whether it is reasonable to expect a high correlation between the SISS Skill Categories, given the structure of the tasks. The allocation of marks for each part of a task is another significant factor. Tamir and Doran (1992b) contend that the skills are hierarchical with "performing" at the lowest level and "reasoning" at the highest level. The initial parts of each task were defined as "performing" — the lowest level in the hierarchy — yet were allocated the major portion of the marks. Many students were successful in completing the "performing" part of a task, but were not able to accomplish the "reasoning" part of the task with similar success. However, those who were successful in "reasoning" were usually successful in "performing" too. Given this hierarchical structure, the low correlations between categories are not surprising.

"Substantial errors of measurement" (Tamir and Doran, 1992b, p. 370) is a catch-all phrase that can be considered to refer to the reliability of the test scores, and its impact upon the validity of score inferences. Tamir and Doran do not articulate whether they believe that such measurement errors are systematic or random. There are many opportunities for systematic errors to occur. These range from lack of precision in construct labelling to problems in translation across different languages. Systematic errors tend to affect scores in a regular or consistent fashion. Random errors derive from purely chance happenings (Crocker and Algina, 1986), for example, guessing, scoring errors or inconsistencies in administration. Any analysis of measurement error requires a careful investigation of the structure, procedures and interpretive framework of the assessment.

The Third International Mathematics and Science Study (TIMSS)

The Third International Mathematics and Science Study (TIMSS) is the current IEA project in mathematics and science. This study has attracted the participation of over 50 educational systems, with extensive representation from countries on all continents except

Africa. The TIMSS performance option attracted the interest of 20 countries and is the focus of discussion here. It must be noted that student performance data and analytical procedures in the TIMSS project are not yet available for analysis.

While many parts of the TIMSS project can be seen as a progression from the Second International Science Study (SISS) and the Second International Math Study (SIMS), there have been significant developments in the design of the TIMSS framework. Most notable is the move away from the content-by-cognitive behaviour grid. Criticism by Romberg and Zarinnia (1987) is identified by Robitaille, Schmidt, Raizen, McKnight, Britton, and Nicol (1993) as having stimulated this change.

> They (Romberg and Zarinnia) note that the use of a content by cognitive behaviour grid fails to take into account the interrelatedness of content or of cognitive behaviours, and that this forces the description of information into unrealistically isolated segments...
>
> The TIMSS curriculum frameworks were constructed to be powerful organizing tools, rich enough to make possible comparative analyses of curriculum and curriculum changes in a wide variety of settings and from a wide variety of curriculum perspectives. The framework had to allow for a given assessment item or proposed instructional activity to be categorized in its full complexity and not reduced to fit a simplistic classification scheme that distorted and impoverished the student experience embedded in the material classified. (Robitaille et al., 1993, p. 42)

The developers of the TIMSS project hope that the criticism of Romberg and Zarinna can be addressed by the use a three-dimensional framework. Three considerations are included: subject matter content, performance expectations, and perspectives or context (see Figure 28).

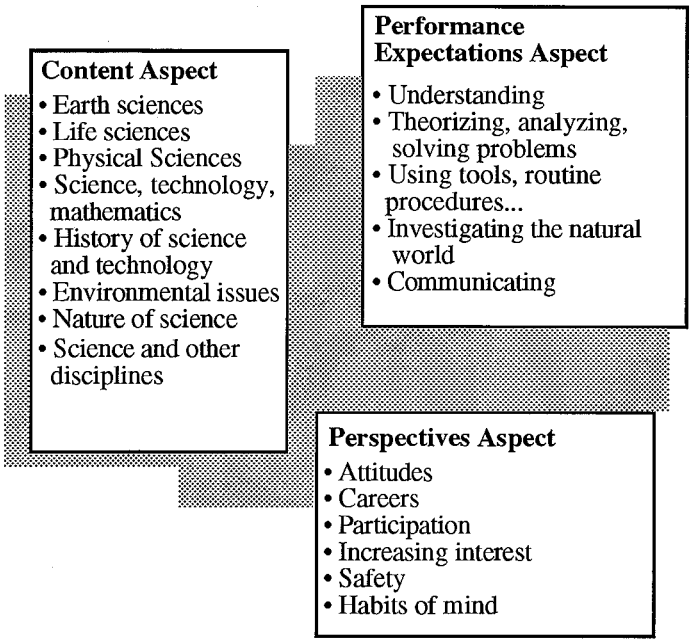


Figure 28. Aspects and Major Categories of the TIMSS Science Frameworks (Robitaille et al., 1993, p. 46)

Robitaille and his co-authors describe the content aspect as "obvious and needs little explanation or development" (1993, p. 44). They regard the performance aspect as:

> a reconceptualization of the former cognitive behaviour dimension. The goal of this aspect is to describe in a non-hierarchical scheme, the many kinds of behaviours that a given test item or block of content may elicit from students.

They state that the perspectives aspect is intended to:

> permit the nature of curriculum components according to the view of the nature of the discipline exemplified in the material, or the content within the material is presented.

The addition of an extra dimension to the TIMSS framework is not the only significant development in the classification of curricula or item analysis. Of major import is the recognition that an item is not necessarily uni-dimensional:

> In the TIMSS frameworks, a test item or block of content can be related to any number of categories within each aspect, and to one or more of the three aspects – thus, the multi-category, multi-aspect designation. It is no longer appropriate to think of disjoint "cells" since hierarchical levels within each category make overlapping cells possible. An item is no longer represented as a single strand linked to a matrix, and instead may be associated with many combinations of aspect categories in the TIMSS frameworks. (Robitaille et al., 1993, p. 44)

Robitaille and his co-authors go on to introduce the idea that an achievement item, or piece of curriculum material, has its own "signature". They expand by saying:

> Technically, a signature is a vector of three components and the three components are themselves each vectors of category codes for one of the three aspects of a TIMSS framework. That is, associated with each item or piece of curriculum is an array of categories from one, two, or all three aspects of the relevant framework...
> The signature reflects the multi-aspect, multi-category nature of the frameworks. It also provides a more realistic depiction of the complex nature of the elements of curriculum, and is less reductionist than the traditional one-to-one mappings. It is more suited to the complexity of student activities emerging from the various national reforms of school mathematics and science, and more suited to the rich, integrated performances expected of students in the new forms of assessment that are emerging along with curricular reforms. (Robitaille et al., 1993, p. 44-5)

In rejecting the simplistic grids of the earlier assessments, the developers of the TIMSS study have provided a structure that is likely to have two significant effects. It will:

1. enable task descriptors to convey the intricate details of complex tasks; and
2. pose a challenge to those who will interpret student scores.

The requirement for uni-dimensional tasks and the problems of describing complex tasks have coalesced into the principal factor limiting the development and use of performance tasks for many years. The developers for both APU and TAPS projects have made significant efforts to create uni-dimensional tasks[6]. Such tasks, particularly those

---

6 Bryce and Robertson (1985, p.18) describe the simplifying of tasks to ensure that they only focus upon one objective as "purifying". Similarly, Johnson (1989) reports that the APU used expert judges to identify a single process from the APU framework for each station task.

developed by the TAPS group in Scotland (Bryce and Robertson, 1985) are limited by this constraint. The "signature" approach proposed for TIMSS enables the development of longer and more complex tasks that are consistent with current reform efforts. However, the use of complex tasks leads to significant problems for those who interpret the scores. The challenge of making sense of the TIMSS scores places the project in a position where it will necessarily catalyze developments in the application of measurement theory of large scale assessments.

Student Performance Component of the 1991 B.C. Science Assessment

The approach taken to describe the tasks in the Student Performance Component of the 1991 B.C. Science Assessment is based upon the premise that content and process are inseparable. A position paper, presented on behalf of the Contract Team to the Ministry of Education at the commencement of the project, clarifies this perspective:

> Although we are concerned primarily with the "skills" dimension in this report, it should be clear that the assessment of these skills can only be accomplished by presupposing that the students possess a related knowledge base and the willingness to articulate that understanding in a given assessment setting.
> ...the descriptors that we will outline below in the framework of students' practical skills in science have been written in as general, content-free terms as possible. This is so that they can be used as a basis for constructing more specific descriptors for the desired age levels as well as the desired content area. (Erickson, 1990, pp. 2-3)

Thus, the framework is written in terms of six dimensions of science, each with a specific subset of abilities (Figure 29, next page). Two types of assessment task were developed: stations and investigations. The investigation tasks are open-ended, in that students plan and perform their own experiments to solve a specified problem. Consequently, the assignment of specific abilities to each investigation is not possible. For the stations, descriptions of observable actions are written in the language of behavioural objectives. The framework was used to describe observable actions of the students, including their

(1) Observation and Classification of Experience

1.a. ability to describe observations made using the senses about a variety of living and non-living objects
1.b. ability to group living and non-living things by observable attributes
1.c. ability to select relevant information from observations to address a problem at hand
1.d. ability to use 'keys' for identifying and classifying objects
1.e. ability to observe and recognize changes or regularities which occur over time
1.f. ability to classify objects in different ways and construct keys to show others how this was done
1.g. ability to observe objects or events from different perspectives

(2) Measurement

2.a. ability to make simple measurements of phenomena using 'invented' or established units
2.b. ability to supply correct SI and metric units for common measurements
2.c. ability to read the scales of various measuring instruments
2.d. ability to estimate the quantity of common properties such as weight and length

(3) Use of Apparatus or Equipment

3.a. ability to identify equipment used in investigations
3.b. ability to state the purpose of a piece of equipment
3.c. ability to select a piece of equipment appropriate for a given investigation
3.d. ability to use a variety of equipment in conducting an investigation
3.e. ability to explain the limitations of specific equipment

(4) Communication

i) Receiving and interpreting information

4.a. ability to follow written, oral or diagrammatic instructions
4.b. ability to draw inferences from data presented in tabular, pictorial or graphic format or data generated experimentally
4.c. ability to translate information from one format to another

ii) Reporting information

4.d. ability to discuss orally the results of an investigation
4.e. ability to report results of an investigation in a descriptive format
4.f. ability to use appropriate symbolic representations when reporting results of an investigation

(5) Planning Experiments

5.a. ability to pose an 'operational question'
5.b. ability to develop a plan that is relevant to an identified problem
5.c. ability to identify the relevant variables that will allow the operational question to be addressed in a proposed experiment
5.d. ability to suggest a testable hypothesis
5.e. ability to identify the possible safety risks in a proposed experiment
5.f. ability to predict the sources of error in an experiment
5.g. ability to use representational models in planning an experiment

(6) Performing Experiments

6.a. ability to set up and use materials and equipment safely
6.b. ability to develop and carry out a suitable measuring strategy for the appropriate variables
6.c. ability to develop a strategy for collecting and recording data that is relevant to the operational question being addressed
6.d. ability to transform and interpret the data collected so that a response can be provided to the operational question
6.e. ability to evaluate the results of the experiment to determine whether further experimentation is required

Figure 29. The Dimensions and Abilities used in the Student Performance Component of the 1991 B.C. Science Assessment

manipulation of equipment and their responses to prompts or questions. It is important to add that each objective is cross referenced to the "Ability" from which it is derived. For example, consider the station "Sound Board", which was used with Grades 4, 7 and 10. The objectives for this station are:

> 1. Students observe and describe what is different about four steel strings of a varied diameter. 1.a
> 2. Students listen to the sound produced by each string and describe the differences they hear. 1.a
> 3. Students explain why the strings sound differently 4.b
>
> (Erickson et al., 1992, p. 20)

These three objectives are considered to encapsulate the key performance elements of the station. The first two are derived from ability 1.a, "ability to describe observations made using the senses about a variety of living and non-living objects". However, each objective is amplified in terms of actual student performance with the equipment, hence the differing statements. The order of the objectives reflects the sequence of instructions or activities as students proceed through a particular station.

How to score students was an issue for the Contract Team in the design of this component of the 1991 B.C. Science Assessment. While this is examined in greater detail in the next chapter, the features related to the description of student performance will be reviewed here. One issue concerns the relationship between stated objectives and how performance is rated on the task. A holistic scoring system was chosen for the Student Performance Component of the 1991 B.C. Science Assessment. One or two carefully considered evaluative or scoring questions — refered to as judgements — were formulated for each station. The questions for the station "Sound Board" are:

> 1. In your judgement how well did the student observe differences between the strings (visual and aural),
> 2. In your judgement how well did the student explain why the strings sound differently. (Erickson et al., 1990, p. 21)

In this case, the evaluative questions are derived directly from the set of objectives, with the first two objectives (both based upon ability 1. a) combined as the basis for the first

question. This is the model for most stations. But for some, mapping of objectives into evaluative questions in this way did not cover all objectives[7]. The Grade 10 station "A Chemical Reaction" demonstrates some of the omissions that can arise when holistic scoring is used. The objectives for this station are:

1. Students measure the pH of water. 2.c
2. Students observe the reaction between calcium and water. 1.a
3. Students measure the pH of the products of the reaction. 2.c
4. Students write a word equation. 4.e
5. Students write a symbol equation. 4.f

(Erickson et al., 1990, p. 171)

These objectives were translated into two judgements:

1. In your judgement how well did the student observe the reaction between calcium and water?
2. In your judgement how well did the student write a symbol equation?

(Erickson et al., 1990, p. 176)

In moving from objectives to evaluative questions, both of the pH measurement objectives were lost to scoring, as was that of writing a word equation. Loss of information about student achievement related to specific objectives will occur if such objectives are not incorporated into the evaluative questions. However, in the Student Performance Component of the 1991 B.C. Science Assessment, descriptive data were collected separately for each student. In this way it is possible to examine individual as well as collective performance in relation to specific objectives.

A MODEL FOR DESCRIBING STUDENT PERFORMANCE

The large-scale assessment programs examined in this chapter have used many different strategies to describe science performance assessment tasks. Achieving a balance

---

7 In order to facilitate the synthesis of the results for the student evaluations, the contract team chose to limit the number of judgements for each station to a maximum of two.

between specificity and generality in definition requires thought and examination of the consequences of decisions made in other large-scale assessments. To be general in categorization, as in the case of the SISS project, leads to descriptors that enable situation of tasks within the domain of science, but allows little specificity regarding task description. Conversely, to be so specific as to describe tasks with minutiae that cannot be adequately addressed with analytical procedures is a strategy that leads to a false sense of precision.

My proposal for task description follows from Shepard's question about what can be claimed from the assessment practice. While hands-on performance assessment in science is very much more than process assessment, it must be recognized that processes or skills are involved. In addition, the content knowledge requirement of each task, and content representation of the domain of the set of tasks, are further elements that must be considered. Each task has a unique identity in terms of the types of practical skills encountered in it, and the content area in which these skills are demonstrated. Yet all belong to the domain called "science".

A model for differentiation between practical skills was developed by Millar (1991). He chooses to distinguish between general cognitive processes and scientific procedures, which he calls "practical techniques" and "inquiry techniques". The critical feature that Millar uses to separate the elements of practical skills is that cognitive processes cannot be taught, but scientific procedures can, and may be improved through teaching. Millar's position extends the representation of procedural understanding developed by Gott and Murphy (1987). In claiming that practical techniques can be taught, Millar asserts that:

> These are specific pieces of know-how about the selection and use of instruments, including measuring instruments, and about how to carry out standard procedures. They can also be seen as progressive, in terms of both the increasing conceptual demand of certain techniques and increasing precision. (1991, p. 51)

The progression from elementary science to secondary science, together with the use of equipment in a school science laboratory, provides strong circumstantial evidence in support of Millar's claims. Teachers instruct students about the specific methods of using apparatus such as mechanical mass-measuring devices[8]. Simple double-bucket balances, with precision to the nearest 10 g used at early elementary levels, are replaced by balances that measure masses to ± 0.5 g at later elementary grades, to ± 0.1 g in early secondary courses and ± 0.001 g in many senior secondary chemistry and physics classes. Not only are senior students working at higher levels of precision, frequently they are asked to analyse the probable sources of error in their equipment or technique. Millar amplifies his characterization of inquiry tactics. These are:

> best thought of as a 'toolkit' of strategies and approaches which can be considered in planning an investigation. These would include repeating measurements and taking an average, tabulating or graphing results in order to see trends and patterns more clearly; considering an investigation in terms of the variables to be altered, measured, controlled; and so on.
> (Millar, 1991, p. 51)

Millar argues that the tactics or strategies of science are best taught through student involvement with investigations, a perspective supported by Woolnough (1991). In examining the status of general cognitive processes, Millar (1991) warns of confusion between "*means* and *ends*. The processes are not the ends or goals of science but the means of attaining those goals" (Emphasis in original; p. 50). Millar's model for the categorization of practical skills is shown as Figure 30.

---

8 This is now the case only for mechanical as opposed to electronic digital read-out balances. The practical techniques for use of electronic balances are generally the same for elementary students measuring masses to the nearest gramme as for senior secondary students weighing to the nearest milligramme.

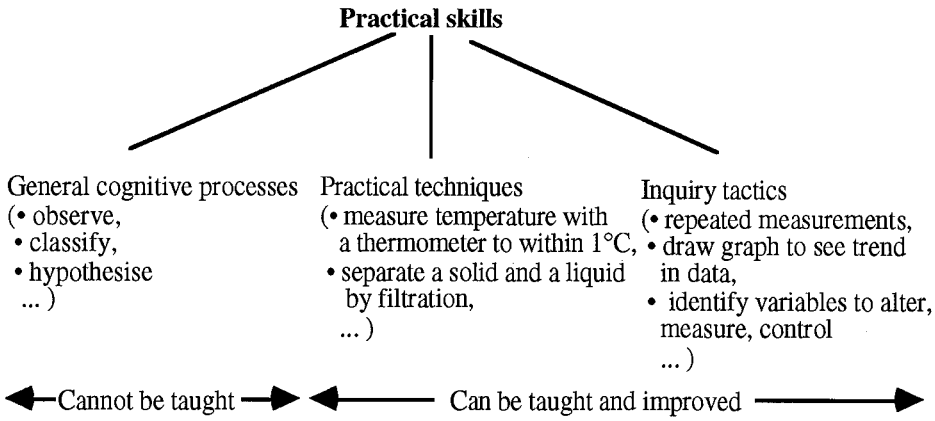


Figure 30. Sub-categories of 'Practical Skills' (Millar, 1991, p. 51)

I have expanded the model proposed by Millar in order to provide a method of describing student performance assessment tasks in science. A fundamental assumption of this perspective is that practical skills must be demonstrated in the context of some science content. In order to escape from the limiting influence of a traditional content by process matrix, I propose a three-dimensional relational diagram, shown as Figure 31.

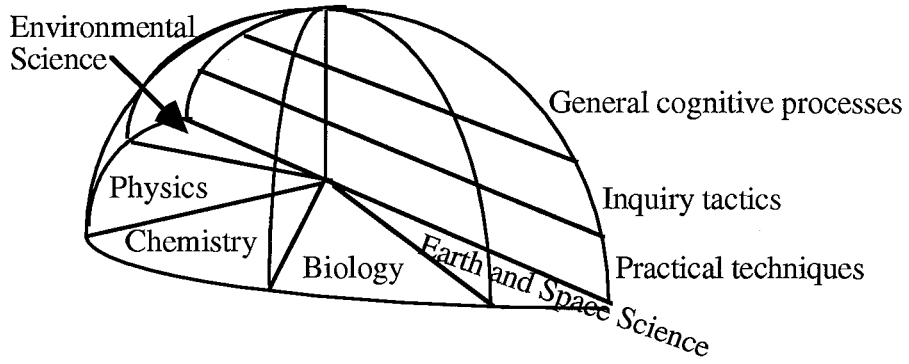


Figure 31 A Three-dimensional Model of Performance in Science

The horizontal plane represents a typical set of content area sectors for school science. The vertical plane shows segments which incorporate Millar's three categories of practical skills. The volume within the quarter-sphere represents the universe of the science curriculum for the assessment. The mapping of individual tasks onto this three-

dimensional model is difficult to show on two-dimensional paper. The simplification of the model into a two-dimensional template facilitates mapping of tasks, but requires a reorientation of the categories of practical skills into sectors of the circle. In the two-dimensional diagram (Figure 32, next page) the area plotted out by a specific task must include both "science content" on the left side and "practical skills" on the right side.

To exemplify use of this template, the station "Environmental Testing" (used with Grades 7 and 10) is mapped onto the template in Figure 32. The objectives presented by the Contract Team for this station are:

1. Students follow instructions.
2. Students make and record observations.
3. Students draw inferences from collected data.

(Erickson et al., 1990, p. 149)

In this station, students are asked to use pH paper to measure the pH of two different samples. One sample is neutral and is labelled "Lake Water 1985" the second is acidic and is labelled "Lake Water 1990". Students are expected to respond to the question "What might have caused this change?" by making inferences about possible causes of the different pH results. The "science content" that pertains comes from chemistry and environmental science, so the left-hand of the template maps to these two sectors. Use of the pH paper is considered a "practical technique", while observation of the colours of the pH paper represents "cognitive processes".

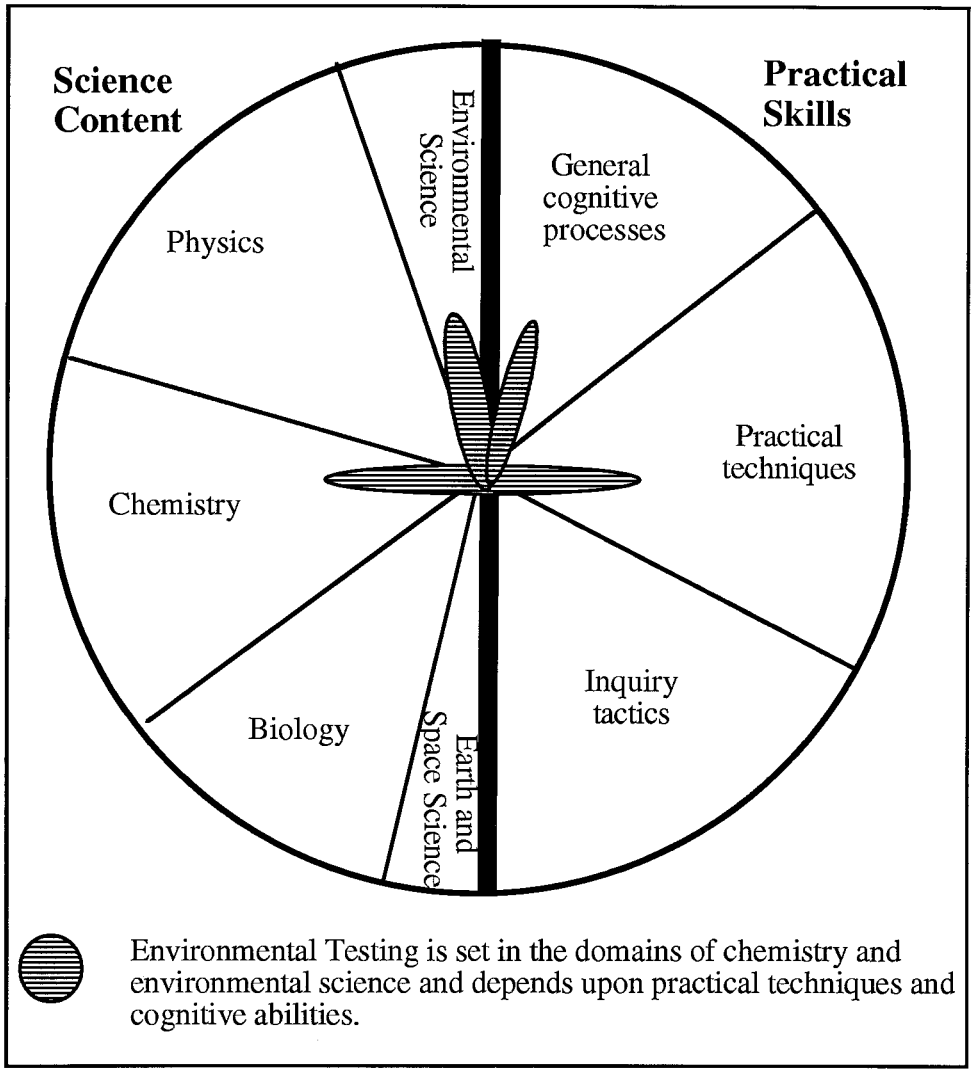


Figure 32. A Two-dimensional Model of Performance in Science

This scheme for mapping perceived attributes of specific tasks onto a relational diagram has the advantage of enabling visualization of individual task specificity and content coverage provided by the set of tasks. Mapping tasks onto this template enables test developers – and users — to be more secure in their claims about what will be or was measured in an assessment. This is illustrated in Figure 33 which represents the Grade 10 stations used in the Student Performance Component of the 1991 B.C. Science Assessment.

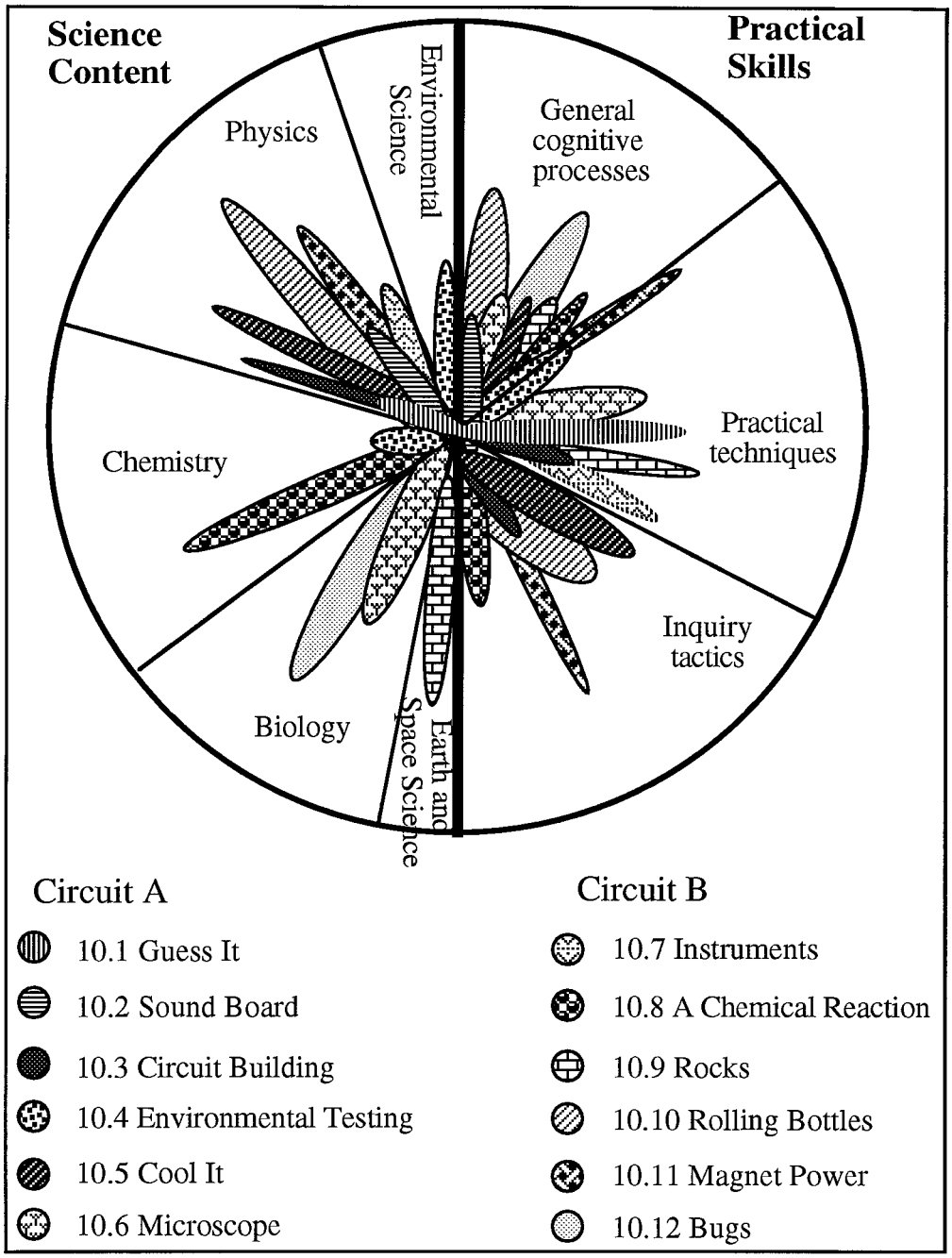


Figure 33. The B.C. Science Assessment Grade 10 Station Tasks Mapped onto the Template (Radial plot dimension does not imply level of content)

Allocation of tasks to sectors of the circle is based upon the analysis of station objectives as written in the *Technical Report* (Erickson et al., 1992). Each of the 12 stations has its own unique pattern. Examination of the distribution of the plots shows that the science content

for most of the stations is in the domain of physics, with a smaller number of stations representing each of the other content areas. Few of the tasks were considered to belong to more than one content area. Conversely, for practical skills, many stations appear to be multi-faceted: each of the sectors is covered by at least six tasks.

This method of drawing tasks on a template also provides a model that helps explain the low levels of across-task generalizability found in performance assessments in science. Discussions of task specificity in measurement journals (e.g., Linn and Burton, 1994; Shavelson et al., 1993) have tended to focus upon the statistical properties of a small number of tasks using generalizability theory (Shavelson and Webb, 1992). The model proposed here can be used to present a set of task structures, with a profile for each task, and the relationship among these profiles can be clearly displayed. An advantage of this approach is that it enables the reader to examine the congruence of tasks and to predict possible degrees of correlation between them.

My vision of the description of performance assessment tasks leads to two distinct approaches that augment each other. One of these starts with the construction of a framework that sets out generic forms of the three sub-categories of Millar's "practical skills". This framework is then used to describe the students' actions in completing tasks, and leads to a set of content-dependent "objectives" (as in the case of the Student Performance Component of the 1991 B.C. Science Assessment) or to a "signature" (as in the case of TIMSS). I recommend that a second stage of description of assessment tasks be built from the objectives or signature developed in the first stage. This involves mapping the tasks' profiles onto a two-dimensional template as shown in Figure 33. If the test developers and test users are satisfied that the domain of the test items provides an adequate representation of the universe of interest, then they can have some confidence in the inferences that they make from student performance on the set of tasks.

CHAPTER 6 — INTERPRETING STUDENT PERFORMANCE IN SCIENCE

This chapter begins with a discussion of the derivation of student scores, and then moves into a discussion of alternatives for interpretation. Once again, construct validity is the driving theme for the questions considered in this chapter — initially in terms of deciding how score meaning is influenced by scoring systems. The procedures for score interpretation are then examined in answer to the third research question of this dissertation:

> What are the implications of using different strategies for scoring student achievement upon the interpretation of student performance?

This leads to the fourth question:

> How could/should test scores in performance assessments be interpreted and used?

An appropriate start is the examination of two extreme approaches to scoring: holistic and analytical. Choice of scoring approach is influenced by a view of the nature of the assessment tasks, and leads to decisions which impact upon how student performance should be evaluated. Having obtained student scores, the critical issues for test developers and users are the establishment of procedures for making inferences about student performance, reporting these inferences, and examining the usefulness of the whole assessment process.

SCORING PROCEDURES

The essential difference between holistic scoring procedures and analytical approaches lies in the process of aggregation. In holistic scoring all information contained within the student response is amalgamated so that a "whole" score on a task is created. Alternatively, in analytical scoring methods, a score is allocated to each element within a

task. Item scores are summed or otherwise manipulated to create a total task score. Crocker and Algina (1986) argue that whenever item scores are summed, the resulting test score should be described as a composite.

Choosing a scoring approach has sometimes stimulated heated debate. Development of the National Curriculum in England and Wales involved particular tension between Woolnough (1989) of England (arguing for holistic assessment), and Bryce and Robertson (1986) of Scotland (favouring an analytical approach). Consider the following section from an article entitled *Practical science assessment. Will it work in England and Wales?*

> ...teachers should no longer be urged to judge or rate 'holistically' their pupils' performances on 'experiments' or 'scientific investigations'. In a comprehensive examination of the international literature we have recently shown that, however desirable, there is no demonstrable evidence of validity and reliability in currently available versions of assessment by holistic teacher-judgement. You may consult the literature yourselves: it is all well documented. Should you do so you will be dismayed by the lack of practicality in the procedures suggested. An alternative to holistic assessment is the use of carefully structured assessment items such as practical test items and 'item-sets' where product checks and end-checks afford the means by which teachers can conduct practical assessment with individual pupils, in classes. (Bryce and Robertson, 1986, p. 63)

Developments in performance assessment techniques, in the years since Bryce and Robertson wrote this article, have addressed many issues pertaining to validity and reliability. These are now reviewed.

Holistic Scoring

The advantage of using holistic scoring is explained by the California Learning Assessment System (CLAS) as that it accommodates "a wide range of student responses, as well as evaluating the student's thinking process" (Comfort, 1994, p. 17). The approach taken in California was to develop a 4-point scale with a generic scoring shell used to provide descriptors for each point on the scale. The descriptors were then used to

derive holistic scoring guides for each task. The scoring-guide shell points 4 and 1 are shown below in Figure 34.

Score Point 4
Demonstrates in-depth understanding of the scientific concept(s) and processes and of the problem and investigations. Responses include all elements of scientific design, and the responses are succinctly and clearly stated. All data are presented in an organized fashion – complete and accurate diagrams, charts, tables, and/or graphs – with clear and effective supporting evidence. All observations are recorded and valid and demonstrate attention to detail. All questions are clearly communicated and ideas are communicated clearly and effectively. Explanations clearly support and demonstrate the relationship between data and conclusions. Responses demonstrate the need for additional testing and provide appropriate suggestions related to the problem (if appropriate for grade level). Scientific terminology is used in context with meaning (if appropriate for grade level).

Score Point 1
Demonstrates extremely limited or no understanding of the scientific concept(s) and processes or understanding of the problem and investigations. Responses include major misconceptions. Responses lack elements of scientific design and no supporting evidence is present. If data are presented they are vague and unorganized. Diagrams, charts, tables, and/or graphs are missing, incomplete, and inadequate. If supporting evidence is present, it is flawed, disjointed or does not match data. Observations are usually missing, and if recorded they are vague with no attention to detail or do not match the task. Attempts to answer a few questions on the task but responses are incomplete, do not show evidence of understanding and may not match the question. Explanations may be missing and student is not able to show the relationship between data and conclusion. Student may rewrite the prompt or write off topic.

Figure 34. CLAS Science Scoring-guide Shell Points 4 and 1 (Comfort, 1994, p. 20)

Comfort (1994) reports that a score of 4 represents performance that is "clear and dynamic", scores of 3 and 2 are characterized as "acceptable, adequate" and a score of 1 indicates performance that is "attempted, inadequate". The California procedure for creating task-specific criteria involves three steps. Student response papers from a field test are sorted into groups representing the four score points. Next, a rationale is given for each grouping. Finally, specific criteria for each score point are drafted with reference to the scoring guide shell. The draft scoring guidelines for score points 4 and 1 of the CLAS task entitled "Spaceship U.S.A." are shown in Figure 35.

Score Point 4	Data are organized in a table/chart with complete and accurate labels. Explanations in questions 11 through 14 show a clear understanding of living/nonliving interdependency. Conclusions are drawn from interpretation of the student's own data on the chart.
Score Point 1	Data are not organized well. Labels are incomplete or inaccurate, and there is no evidence of conceptual understanding.

Figure 35. CLAS Science Score Points 4 and 1 for "Spaceship U.S.A" (Comfort, 1994, p. 32)

Once all score points have been described, "anchor papers" are selected to "represent a range of high, medium and low student achievement within each score point" (Comfort, 1994, p. 21). The score points established by this process are used at all scoring sites across the state. The scoring procedures for the CLAS science assessment are reported to have acceptable levels of inter-rater reliability, with coefficients typically greater than 0.8 (Comfort, 1994). An extensive set of operations lead selected teachers to work through tasks, review scoring guides, and discuss specifications for the assessment task with pre-trained table-leaders, all under the watchful eye of "chief readers" who were involved in defining the score points. After this orientation, the teachers worked on scoring a range of anchor papers. Eventually they arrived at what was called the "calibration round" in which they all read the same anchor paper and defended the score they assigned. This process was intended to ensure consistency.

> In order to score live papers, readers must qualify during this calibration round. If a reader strays from the scoring guide while scoring live papers, the table leader is responsible for realigning the reader to the scoring guide. (Comfort, 1994, p. 32)

The number of student scripts in California is immense: over 400,000 fifth grade students were tested in 1994 with six sites designated as scoring centres. Consequently, the training and calibration procedures used were intended to be overtly structured, with the hope of ensuring reliable scoring at the level of the individual student.

The CLAS science assessment designers set out to develop long "coordinated" tasks. For example, "Spaceship USA" (Comfort, 1994) saw students respond to 14 questions in a period of 40 minutes. Comfort reports that "Spaceship USA" was assigned a single holistic score in the 1991 field trials. When it was used for further field trials in 1993, "Spaceship USA" was scored using logical groups of questions in order to provide a set of component scores (Comfort, 1994).

The scoring procedures for the Student Performance Component of the 1991 B.C. Science Assessment were also holistic in orientation, but the administration of scoring procedures was not weighed down by the massive numbers of the Californian sample. The procedures for developing scoring criteria in the B.C. project are set out in Chapter 3 of this dissertation. Of particular significance is the Contract Team's decision to define the key aspects of each task in the form of one or two specific questions, but to delegate the definition of criteria for specific levels of achievement to the teachers who administered the assessment. The Contract Team argues that this approach enabled the teachers to make:

> a global judgement based on their professional experiences of students at that grade level. They were not asked to think of the task other than in its entirety, as a procedure that related to the actions of the students as they would perform them in every day life. It was this type of holistic judgement, related to the real life and real time aspects of the student's actions, that we felt the teachers were best able to perform, given their previous experience with the tasks and with students in general. (Erickson et al., 1992, p. 252)

The effects on the validity of the assessment process of enabling teachers to derive the scoring criteria are discussed in Chapter 4 of this dissertation.

Analytical Scoring

The perceived advantages of analytical scoring procedures tend to revolve around the fact that a greater number of score points can be generated to demonstrate student achievement in specific parts of a task. Item scores are usually aggregated to produce a

composite score. Within an analytical scoring framework it is important to direct attention towards the nature of the item scores. Are scores to represent a continuum of achievement? Are they to be considered in terms of mastery of specific objectives? In my view, the nature of particular assessment tasks should guide the structure of the scoring procedures. Anastasi (1988) proposes that pass/fail scoring be limited to tests of mastery of basic skills. The development of the program "Techniques for the Assessment of Practical Skills in the Foundation Sciences" (TAPS) (Bryce et al., 1984) identified as a first phase the development of items to assess basic skills in science. TAPS was developed with the expectation that once such skills were mastered, students would move on to "more complex processes and strategies of scientific method" (Bryce and Robertson, 1986, p. 63). The TAPS tasks have been criticized on the grounds that the nature of science presented by the tasks is overly simplistic and should not be represented in this way (Millar and Driver, 1987; Woolnough, 1991). Hanna (1993), writing about mastery tests in general, identifies the core of the problem:

> Assessing all-or-none attainment of rigidly sequenced objectives is unnecessary when (a) the content is not inherently sequential and (b) achievement is not dichotomous. (p. 64)

A mastery based scoring and reporting system was used in the most recent international assessment, the 1991 International Assessment of Educational Progress (IAEP). The report, *Performance Assessment: An International Experiment* (Semple, 1992), presents the scoring scheme in terms of "credit for..." and gives results for students from different jurisdictions in terms of "Percentage of correct responses". The percentage of students succeeding ranged from 10% on certain tasks up to 100% on others. An example of the style of data presentation is shown in Figure 36, although the data in this figure do not represent actual results.

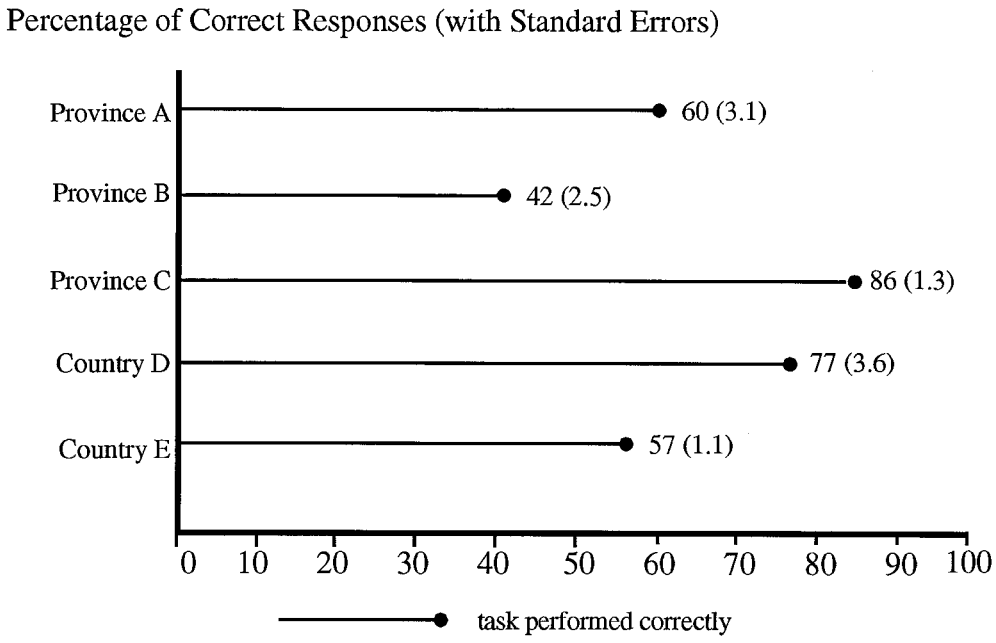


Figure 36. IEP Style of Presenting Scores (after Semple, 1992)

In Semple's report there are fewer than three pages about interpretation of the numbers, and the procedures used for interpretation. This absence may justify the criticism that "the performance-based items were not of the same quality as the multiple-choice questions" (Rothman, 1990b, p. 10). This was the reason given to excuse the absence of the United States from this project despite the fact that it was coordinated by a division of the Educational Testing Service in Princeton, New Jersey!

Hardy reports on a study in which four science performance assessment tasks were designed and administered to elementary school students in Georgia (1992). Data are presented for three of the four tasks that were scored using both analytical and holistic scoring rubrics. The time needed to score the tasks was found to be similar for each rubric and was "directly related to the complexity of the task, e.g., the type and number of judgements required for each student score". Hardy goes on to comment that inter-rater reliability:

> ...ranged from a low of 0.76 for holistic scoring of *Design a Carton* to no discrepancies in scoring (r = 1.00) for the objective scoring of *Discovering Shadows* and the holistic scoring of *Shades of Color*. (1992, p. 18)

Other values for inter-rater reliability coefficients are greater than 0.85. The correlations between analytical and holistic scoring are also described:

> The correlations of analytical and holistic scores for *Shades of Color* and for *Identifying Minerals* were r = 0.66 and r= 0.78 respectively. Correlations of this magnitude are common for comparisons of analytic and holistic scoring for essays and writing tasks. The correlation of analytical and holistic scoring for *Design a Carton* was r = 0.33. This relatively low correlation suggests that, for this assessment, the two methods of scoring are likely measuring different constructs. (Hardy, 1992, p. 18)

The data for the task *Design a Carton* led Hardy to comment about how the design of the scoring rubric affected student scores:

> An inspection of the data distributions suggests that the difficulty of the task for the target population produced a restriction of range that was more evident in the holistic scoring than in the analytical scoring. Although score points of 0-5 were anticipated for the holistic scoring, no scores of 4 or 5 were awarded. These points were possible only if a student group considered more than one carton design and showed a comparison. The restriction in range was less evident in the analytic scoring. Consequently, score points of "2" or "3" by the holistic rubric were correctly associated with a range of score points by the analytic rubric. Conversely, papers assigned a "3," "4," or "5" by the analytic rubric are all classified as "2" by the holistic rubric. These differences point to the difficulty of developing rubrics when only limited data are available for tryout. A modification of the holistic rubric to better differentiate the actual range of responses would likely lead to a higher correlation with the analytic scoring. (1992, p. 23)

The issues that arise from choice of scoring approach that artificially restrict the range of possible scores must be addressed in the design of assessment tasks, and in the piloting of scoring procedures. The problem appears to surface frequently and requires careful attention. For example, the Contract Team for the Student Performance Component of the 1991 B.C. Science Assessment observed that:

> The teachers' judgements as to how well the students observed differences among the beans is puzzling. Sixty-one percent of students were judged to have performed at a satisfactory or better level, while only 1% were judged to have performed "extremely well." Teachers judged student performance by the number and complexity of the criteria they used. Category 3 required 1 criterion, category 4 required 2 criteria while Category 5 required 3 or more criteria. We think that "more is better" is a teacher judgement which disadvantages students. Some students may well value one salient criterion as an appropriate way to observe and designate similarities. It is in this adult perspective of appropriateness that the student performance may be undervalued. (Erickson et al., 1992, p. 19)

While the original design of Student Performance Component of the 1991 B.C. Science Assessment used holistic scoring, a consultant in one school district has recently rewritten the scoring guides for the Grades 7 and 10 stations (Klinger, 1994). The revised forms use analytical scoring techniques. The provision of a "marking scheme", rather than holistic descriptors for levels of performance, is seen by Klinger as more analogous to the regular practice of secondary science teachers.

A challenge for developers and users of assessments is to go beyond the explicit statements of criteria and to reflect upon the type of performance that deserves a particular score. Although the criteria for categories of performance were drawn up by practicing teachers who themselves had experienced both working through the task and administering the assessment, there appears to be little congruence between expected and actual levels of student performance. Other teachers, who were involved with pilot-testing the package *Science Program Assessment Through Performance Assessment Tasks: A Guide for School Districts* (Bartley et al., 1993) described some of these criteria as "blatantly unfair" and argued that the criteria should be revised to concur with what the students actually were asked to do.

Scoring procedures need to be sensitive to the context of the task, as well as able to differentiate between performance levels. Analytical approaches have significant value when the focus of the assessment task is either mastery of clearly specified skills or recall of facts. For open-ended problems, holistic scoring rubrics offer opportunities to evaluate

a range of different approaches to solving the problem. If, as Hanna (1993) argues, the purpose of educational measurement is to inform teaching and learning, then carefully crafted holistic scoring rubrics have great potential for educating students about the types of performance valued by their teachers. Analytical approaches have value in the identification of specific, well-defined competencies, and should also have a place in any assessment structure.

PROCEDURES FOR INTERPRETATION

The final piece of the construct validity mosaic is the interpretation of the assessment data. Implied here is also an evaluation of the assessment process itself. The work of Kaplan (1964), Messick (1989a; 1989b) and Cherryholmes (1988) points to the importance of considering the values and expectations of those persons empowered to interpret test scores. Berlak (1992) believes that the question of assessment context must also be considered. He goes so far as to suggest that an emerging contextual paradigm should be preferred to a traditional psychometric paradigm. Warning against the use of dichotomous categorization of the paradigms, Berlak identifies four foundation assumptions of the psychometric paradigm and presents counter assumptions. My summaries of his analysis are shown in Figure 37 (next page).

Berlak and his co-authors of *Toward a New Science of Educational Testing and Measurement* (Berlak, Newmann, Adams, Archbald, Burgess, Raven, & Romberg, 1992) pose many challenges to those involved in educational measurement. Two crucial factors make their critique of traditional measurement practices particularly influential. First, alternative forms of assessment have gained favour with many educational policy-makers, teachers and students. Second, Berlak's group has access to a large audience through articles in journals such as *Educational Leadership* and *Phi Delta Kappan*. Romberg's work (with Zarinnia, 1987) has been recognized by the authors of the TIMSS framework

Psychometric Paradigm	Contextual Paradigm
Assumption 1: Universality of Meaning. There exists "universality of meaning" where there exists a single meaning about exactly what a standardized or criterion referenced test is claimed to measure. Such agreement in meaning transcends social context and history.	*The Counter Assumption: Plural and Contradictory Meanings.* Plurality of meanings, differences and contradiction in perspectives are inevitable in a multi-cultural world where individuals and groups have different histories, divergent interests and concerns. Validating educational tests based on psychometric canons represents a quest for certainty and consensus where certainty is impossible, and agreement is unlikely unless differences are suppressed and consensus is overtly or covertly imposed.
Assumption 2: The Separability of Ends and Means and Moral Neutrality of Technique. If tests are constructed and interpreted according to accepted standards, they are scientific instruments which are value neutral and capable of being judged solely on scientific merits. Standardized and criterion-referenced tests represent an advance over prescientific and subjective forms of assessment. The ends and the means of testing are seen to be separable. The assessment scientist is best equipped to make judgements about means, that is to develop the ways of assessing educational outcomes and how these are to be properly used and interpreted.	*The Counter Assumption: The Inseparability of Means and Ends.* The argument as to whether a test measures what it is claimed to measure rests upon the case made for its construct validity. Judgement about a test's validity requires choices among contradictory values, beliefs and schooling practices. If judgements about assessment procedures and testing are left to experts, then they assume the responsibility for resolving differences over basic moral questions which in a democratic society should be settled by ordinary citizens and/or their democratically elected representatives.
Assumption 3: The Separability of Cognitive from Affective Learning. The assessment of learning outcomes must be separated into distinct and mutually exclusive categories, separating cognition or academic learning from affect, interests, or attitudes. There exists a three-way classification of human learning which divides head, hand and heart, i.e., the realm of the intellect, of feelings and values, and of manual dexterity. Bloom's *Taxonomy of Educational Objectives* (1956) legitimated the distinctions that are now treated as virtually self-evident in the discourse of teachers.	*The Counter Assumption: The Inseparability of Cognitive, Affective and Conative Learning.* Raven (1992) disagrees with this classification which separates the cognitive from the affective, but also subsumes the conative under "affective." Conative behaviour is that related to determination, persistence and will. Development of human capacities is highly contingent upon the social context, as well as the learner's will, interest and knowledge.
Assumption 4: The Need for Control from the Center. Testing and assessment procedures are a form of surveillance whose use is a super-imposition of a power relationship. Tests shape the school's curriculum and pedagogy – a form of social control. The technology used in the assessment process will encourage particular forms of management and human relationship with the organization, while suppressing others. Mass administered tests are suited to exercising control from the center, e.g., district office, and objectify the subject by reducing all human characteristics to a single number.	*The Counter Assumption: Assessment for Democratic Management Requires Dispersed Control.* There is a need to change the uni-directional nature of the power relationship that has been imposed by the use of standardized and criterion referenced tests. There must be *re*-form of the system of assessment in such a way that it disperses power, vesting it not only in administrative hands but also in the hands of teachers, students, parents and citizens of the community a particular school serves. It is clear that good schools require a strong measure of autonomy by teachers, other school-based professionals, and participation by the local school-community.

Figure 37. Berlak's Measurement Paradigms (1992)

(Robitaille et al., 1993) as a significant factor in their decision to move away from the traditional grid. The impact of *Toward a New Science of Educational Testing and Measurement* is such that the National Council of Measurement in Education (NCME) reviewed it in both association journals. The review in the *Journal of Educational Measurement* was written by Loyd (1994), and that in *Educational Measurement: Issues and Practice* was done by Lane (1994).

Berlak (1992) identifies the key opportunity in a contextual paradigm as holding the power to make decisions that affect oneself. Cherryholmes (1988) is similarly direct:

> Construct validity requires occasional departure from conventional practices of the past, or else the past will dominate the future. Construing construct validity as radical, critical pragmatism takes Cronbach and Meehl another step in the direction they headed in 1955.
> Construct validity and validation is what those in authority choose to call it. If they choose to exclude phenomenological investigation, so be it. If they choose to exclude ideological criticism and discussions of power, so be it. If they choose to stipulate meanings for constructs and enforce them, so be it. But embracing what is in place does not free authorities from the strictures of inherited discourses. They simply are indulged for being in a privileged position. At bottom, construct validation is bound up with questions such as these: What justifies our theoretical constructs and their measurement? What kinds of communities are we building and how are we building them? (p. 129)

I agree that alternative voices should be included in the interpretation process. This leads me to elaborate on specific areas of concern that should be considered in the interpretation of hands-on performance assessment tasks in science. My intent is to clarify the features that affect what can be claimed as a result of the administration of an assessment.

INTERPRETATION OF HANDS-ON PERFORMANCE ASSESSMENT IN SCIENCE

A series of structural considerations for the interpretation of performance assessments in science is proposed in this section. These are articulated in terms of a set of

pre-suppositions, followed by specific concerns related to each part of the process[1]. While the aspirations of Berlak and his co-authors (1992) lead them to argue for a contextual paradigm, authors such as Cronbach (1988, 1989), Messick (1989a, 1989b, 1994), Shepard (1993) and Moss (1992, 1994) have suggested a broadening of the interpretation process as part of the expansion of construct validity. The first set of pre-suppositions arises from considerations pertaining to the purpose of testing, the second set relates to the nature of assessment tasks. The third set is intended to generate discussion about the attributes of the personnel who score and interpret student responses.

Those who develop assessments are responsible for the identification of the purpose of the assessment (AERA, APA, & NCME, 1985). In defining the statement of purpose for hands-on performance assessment in science, I believe four areas should be articulated:

1. A set of explicit goals for the assessment, together with an analysis of how the goals have been operationalized (see Chapter 4 of this dissertation).
2. The intended student population for the test, with details of pilot testing procedures that were used in test development for this population (see Chapter 3 of this dissertation).
3. Purposes and populations for which the set of test items should **not** be used, together with an extended justification of such claims.
4. Intended changes that the testing program is intended to effect with an analysis of the nature of potential unintended consequences (see Chapter 4 of this dissertation).

These four statements are directed to the assessment developer who must specify the intended purpose for assessment tasks, and also must identify applications for which the set of tasks are inappropriate, and state why this is so.

---

1 It is not the author's intent to replace the *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 1985), but to identify specific features that may impact upon the validity of the inferences that might be made on the basis of student scores.

My second area of concern is the nature of school science that is portrayed by the assessment tasks. Again, the onus is upon the developer to explain:

1. The curriculum emphases that were seen to drive the development of the assessment tasks.
2. The perspective towards learning that guided the development of the assessment items, and how such a predisposition influenced item development.
3. The rationale for the types of task and their selection.
4. The domain represented by the set of test items and its representation of the universe of interest.
5. An articulation of the proposed system of scoring.

These five areas reflect the requirement for the developer to identify and defend the assumptions that influence the development of the assessment instrument.

The final important area I shall discuss is the question of who should interpret test scores. Pertinent to this issue are two statements in the *Standards* (AERA, APA, & NCME, 1985).

> **Standard 5.5**: Test manuals should identify any special qualifications that are required to administer a test and to interpret it properly. Statements of user qualification should identify the specific training, certification, or experience needed. (*Primary*)
>
> **Standard 8.2**: Those responsible for school testing programs should ensure that the individuals who use the test scores within the school context are properly instructed in the appropriate methods for interpreting test scores. (*Primary*)

These statements were written in 1985 and are undergoing revision; drafts of new standards are expected in 1995. Application of Standard 5.5 as written above should restrict the interpretation of most assessments to those few experts with specific qualifications in educational measurement. For hands-on performance assessment in science, there are other factors which are worthy of consideration and could broaden the membership of a potential interpretation panel.

1. There are measurement personnel whose training qualifies them to participate in the process. However, if they have not worked through the tasks, or administered the assessment to students, or graded the response sheets, these people may be eminently qualified but their input needs to be augmented as they lack appropriate science curriculum experience.

2. Test developers explore the nuances of each task and are very much aware of the measurement properties of items and how students are likely to respond. Unless this group includes practicing teachers, its members will not be aware of current classroom practices and therefore will have insufficient information and/or experience to make valid interpretations of scores without the participation of others.

3. Science educators from university faculties of education are able to provide a considered perspective about student learning in science. However, unless such personnel have extensive experience in the design of performance assessment tasks, and the specific set of tasks for consideration, useful input is likely to be limited. However, such people are valuable members of an interpretation team.

4. Teachers who have been trained to administer the assessment are constructive interpreters as these people have themselves completed the tasks, and assigned scores to the student responses. Their professional classroom experiences enable them to be aware of the relationship between the test content and the actual school curriculum. It is unlikely that they will have undergone extensive training concerning the issues of test interpretation, but they are valuable contributors to the interpretation team.

5. Politicians, school trustees, parents and representatives of the community are stakeholders whose everyday existence does not take them into classrooms. Nevertheless, their interest in student performance is evident, as is their experience and expertise.

6. Students are increasingly taking greater responsibility for their own learning. A logical consequence of this development is the inclusion of students in the interpretation process. While students lack the experiential background of the other potential members of an interpretation panel, their participation provides an opportunity to expose other members of the education community to students' unique perspectives towards the process.

The package for district assessment, *Science Program Assessment Through Performance Assessment Tasks: A Guide for School Districts* (Bartley et al., 1993) identifies a preferred mix of people for the interpretation process:

> Those teachers who administered and coded the assessment are essential members of the panel. Their knowledge of the purposes of each station, the coding and judgement process, and their sense of how well students did and should do will be invaluable. To provide a different and a wider perspective it is suggested that other teachers, school and district administrators, and trustees would constitute a representative and powerful forum for interpretation. As suggested earlier, the constitution of the panel should be determined in the planning stages of the assessment process to allow for all members to participate in orientation workshops.
>
> It is recommended that the interpretation panel work in grade level teams. Each team would consist of between three to eight people representing the various special interests identified earlier. The presence of teacher/administrators in each grade level group is essential. (Bartley et al., 1993, p. 35)

The author's experience in developing this district package, and a subsequent series of workshops around British Columbia, was that teachers who first work through performance assessment tasks, and then observe students from their own classes under assessment conditions, are pleasantly surprised by their students' performances.

Students can be included in the interpretation process in many ways. This might start at the level of discussions about students' impressions of the tasks during piloting, or could involve teachers discussing the tasks with students after the assessment has been administered. Alternatively, teachers could collect written responses at the time the assessment is completed. To involve students in the grading process offers an exceptional

opportunity for teachers to discuss what they value in student performance, and this would probably lead to significant student growth. However, tensions might appear with the proposal that students be included in the interpretation panels, and this approach might not be acceptable in every political climate.

It is crucial that members of the interpretation panels be aware of the nature of the tasks. The most appropriate route for enabling participants to achieve this state of awareness is to require them to work through the tasks **under actual assessment conditions**. This approach has been successful in many jurisdictions that have introduced performance assessment tasks. The contextual element is also a vital consideration. It is extremely important that classroom teachers participate, in order that their awareness of how specific tasks are synchronized with district or classroom practices be included in the interpretation process.

CHAPTER 7 — CONCLUSIONS

The purpose of this study was to develop a framework for the validation of the use of hands-on performance assessment tasks in school science. The first section of this chapter is composed of my responses to each of the four research questions. In the next section I draw together the claims that I have made in responding to the questions. A discussion about the use of performance assessment tasks in science is included in this chapter which concludes with a statement of the implications for future research.

The systematic framework that I propose in response to the first question provides direction for the validation inquiry. However, in order to evaluate the overall utility of the assessment process I strongly believe that it is vital to consider the complete set of research questions. This is because the values of those who propose, develop and use an assessment must be illuminated.

RESPONSE TO THE FOUR RESEARCH QUESTIONS

1. What are the essential components of a systematic framework for the development and administration of performance assessments in science?

Analysis of the requirements for validation serves to shape the framework that I advance in response to this question (see Chapter 4). My view of the process of validation has been guided by the perspectives towards validity elaborated by both Messick and Shepard. Messick writes of the need to consider "functional worth" (1989b, p. 17) while Shepard poses her validity question in the form "What does the testing practice claim to do?" (1993, p. 429). My interpretation of these authors' work leads me to conclude that a framework for the development and administration of performance assessments should provide an appropriate foundation upon which the final inferences can be built, and in relation to

which the functional effects of the testing process can be analysed. The framework presented in Chapter 4 contains eight headings which identify areas of to be considered in the collection of evidence. These are:

(1) Purposes of the Assessment

(2) Learning and Communication in Science

(3) Content Analysis

(4) Instrumental Stability

(5) Administration Stability

(6) Internal Consistency and Generalizability

(7) Fairness

(8) Consequences

This approach represents a significant expansion in the evidential basis for a comprehensive and unified analysis of the validity of assessment instruments.

The examination of purposes and consequences of an assessment represents much more than beginning and end points. While the "grand purpose" usually remains consistent throughout the use of a particular assessment instrument, the operational purpose demands careful attention during development. As the assessment proceeds, opportunities abound for subtle deviations from the explicit purpose. That purpose and consequence should be inter-connected is self-evident in the analysis of intended consequences, but the search for unintended consequences is also an essential element of this, and every other type of assessment.

The fact that a model of learning and communication in science is conveyed by assessment tasks is a consequence that requires special attention. The impact of basic skills testing on what teachers present within their classrooms and what students learn is described in Chapter 1. Baron (1991) writes of the "magnetic attraction" of tests, and of how good assessment serves instruction well. In the language of psychometrics, the

domain of a test should provide an adequate representation of the universe of interest. A test which appears to require the student to look for a single correct answer does not serve as an adequate representation of science. However, the type of science espoused in the *Draft of the National Science Education Standards* (National Research Council, 1994) has been characterized by the science assessment standards working group as "authentic" in approximating how scientists do their work.

Content analysis and the monitoring of instrumental and administrative stability require that some attention be given to the more traditional approaches of evaluating validity. This is not any less important in performance assessment in science, but it is not the ultimate goal. Instead, these considerations take their place alongside other worthy validity concerns.

My intent in developing this framework was to ensure that the process of assessment is defensible, that claims about student performance are justifiable, and that there is vigilance for unexpected consequences. In order that a model be of practical use, its application in real life must be feasible. The framework proposed has been applied in an evaluation of the development and administration of the Student Performance Component of the 1991 B.C. Science Assessment, where its utility is demonstrable.

2. *What are the essential characteristics of descriptors of student performance on performance tasks in science?*

A starting point in the description of practical skills in science is provided by the work of Millar (1991). I have proposed a further refinement to illustrate the context in which these practical skills are demonstrated. The question "what does this task measure?" is better rephrased as two questions: "what can be reasonably claimed that this task measures?" and "what do students do as they work through this task?". Complex performance assessment tasks tend to be difficult to describe since they are multi-dimensional, including

requirements in both skill and content knowledge. The model I propose facilitates examination of the interaction among practical skills and science content. For example, the skills and content demands of a single task "Environmental Testing" are illustrated in Figure 32. In Figure 33 I demonstrate the utility of the model in presenting the complete set of stations used at Grade 10 for the Student Performance Component of the 1991 B.C. Science Assessment. A plot of the stations upon the template clearly portrays the science content and skills to be assessed by the set of tasks. There are two major uses for such a plot: first to illustrate the adequacy of content representation, and second to analyze task overlap in science content and skills. Considerable variation in student performance across different tasks has been a consistent finding in performance assessments in science (Shavelson et al., 1994). Theoretically, a student should perform at a similar level on tasks which occupy overlapping areas on the template. This expectation is based upon the premise that the tasks measure similar attributes.

A further element in my description of hands-on science assessment tasks comes in the form of statements about the students' actions, rephrased as objectives for each task. The behaviorist elements of describing what students do cannot be denied, but the narration of student behaviour takes the description of the task many stages beyond the provision of an equipment list and a set of instructions. The most suitable manner of conveying the attributes of a task is to administer the task to the interested parties, and then enable these people to observe students attempting the tasks under assessment conditions.

3. *What are the implications of using different strategies for scoring student achievement upon the interpretation of student performance?*

Complex performance assessment tasks may suffer greatly or gain dramatically by the division of each task into component skills in order to develop scoring rubrics. Analytical scoring systems are most promising where there is a need to focus upon specific

responses, for example, in a task to measure specific lengths. The holistic approach to scoring is generally less prescriptive. It has demonstrable advantages when tasks offer students latitude in performance, for example, a "design an experiment to..." type of task. I conclude that the development of a system for scoring performance assessment tasks requires careful analysis of both the nature of the task and the scoring system.

4. How could/should test scores in performance assessments be interpreted and used?

Who should interpret student performance assessments in science is a question of major importance. The constituency of interested parties is large, although some have limited experience or qualifications. I believe that those familiar with the educational context of the curriculum and classroom, and those with knowledge of the measurement issues, should have the most input, but others such as parents, students and members of the community should be included to ensure that the procedures for interpretation are fair and as free from bias as possible.

SUMMARY

In responding to the four research questions, I have described the critical features of a framework for validation inquiry into the use of hands-on performance assessment tasks in the curriculum area of school science. My decision to propose and use a validation framework that features an "integrated evaluative judgement" (Messick, 1989b, p. 13) rather than a group of loosely linked investigations, sets this work within current measurement theory. Developments in conceptualizations of validity, and consequent new directions for validation inquiry, have had a major impact upon the design of the framework and the approach taken to interpret validity evidence. I must emphasize that responses to measurement-driven validity questions are entwined with the theories of learning science that were used to develop an assessment framework. School science is the

focus of my discussion of learning theories. Permeating the whole analysis of every assessment strategy is an underlying theory of learning. Shepard (1991) is critical of measurement-driven instruction because of its underlying model of learning:

> Bracey, Shepard, and others disagree fundamentally with measurement-driven basic skills instruction because it is based upon a model of learning which holds that basic skills be mastered before going on to higher order problems, as Popham has suggested when he says, "Creative teachers can *efficiently* promote mastery of content-to-be-tested and then get on with other classroom pursuits" (1987, p. 682). (Shepard, 1991, pp. 2-3)

This debate within the measurement community parallels the disagreement between science educators who argue for the "step-up approach", for example Bryce and Robertson (1986), and those who propose a holistic approach to assessments such as Woolnough (1989) or Millar (1989).

Performance assessments have generated much discussion in the measurement literature as theoreticians attempt to explain the measurement properties of items where students are offered considerable latitude in their responses. Classical test theory has been found wanting in this regard because of task sampling variability and the low number of items usually included. Generalizability theory has been used by Shavelson and his associates to examine the characteristics of performance tasks in science (Shavelson et al., 1994). They have had mixed success in predicting the number of items required to produce an acceptably generalizable score. The difficulties that these authors found were, I believe, a direct result of their failure to examine the problem from the perspective of learning in science.

The challenge for those involved in examining the technical adequacy of performance assessment instruments is to include the cognitive components of the items, and to consider the characteristics of the population used to generate the data, in addition to examining the statistical properties of instruments and data. With this in mind, I conclude that:

> The development of performance assessment tasks in science requires a clear enunciation of the perspective towards student learning that has driven the assessment.

I hasten to add that merely to identify a learning theory used in an assessment design is insufficient. Test developers must be able to clarify the assumptions used in operationalizing the theory in the context of the assessment tasks. There are several important constraints. These are:

1. The developer must use a theory of learning science to justify task structures.
2. The questions asked and decisions made during pilot studies and field testing should follow directly from the stated theory of learning.
3. The scoring system should reward features of performance that are consistent with the theory of learning.
4. The theory of learning should be used to predict and explain correlational evidence of convergent and divergent validity.

The choice of a theory of learning in which to embed the design of an assessment system is a defining point for that assessment, and must be clearly articulated for those who will use the items and interpret the performance. This consideration should be examined as part of the validation framework. The importance of clarifying the interpretation of the learning perspective is shown in the following example. The constructivist perspective towards learning in science was identified as inspiring the design of two of the assessment projects described in this dissertation. The Contract Team for the Student Performance Component of the 1991 B.C. Science Assessment distinguished a "basic tenet of constructivism that the child is both capable and indeed must experiment in order to construct meaning" (Erickson, 1990, p. 6). This led to a structure with many open-ended tasks. Conversely, reflecting upon her experience as research officer for the TAPS project, Robertson states that the "TAPS research was strongly influenced by the constructivist approach, based upon the findings of Piaget and Bruner" (1994, p. 1). The TAPS group perceived that the work of

Piaget and Bruner substantiated the design of a project for classroom assessment of science process skills; the resultant series of assessment tasks is highly structured with many closed questions. Both these projects claim to be constructivist in orientation, yet the range of tasks and the scoring systems are too different to be explained solely by contextual features. The reason for highlighting such striking differences in design is to show that it is insufficient merely to state a view of learning. Developers must clearly enunciate specific details of the chosen perspective, review the implications of its choice, and carefully apply this perspective in attempting to make sense of student performance.

FINAL REMARKS

Much discussion in the measurement community was generated by Messick's (1989b) treatise on validity. Unfortunately, instances of the application of a unified view have been limited. Indeed, both of two recent validation inquiries[1] are framed with traditional views of validity evidence. Similarly, the 1993 measurement text by Hanna identifies Messick's unified approach as "an alternative view" while the traditional separation of types of validity evidence is described as "common usage" (1993, p. 408). Application of the unified view may be hampered if the examination of consequential evidence appears to be never-ending (Shepard, 1993). A further philosophical issue must be considered in the examination of the underlying values in assessment design and validation inquiry. To follow the guidelines set out in the *Standards* (AERA, APA, & NCME, 1985) is to accept values of the measurement community articulated over 10 years ago. The clarification of values in assessment has been a consistent part of Messick's

---

1 Both studies appeared in *Educational Measurement: Issues and Practice*. The titles are "Determining the Validity of Performance Based Assessment" by Burger and Burger (1994) and "The Vermont Portfolio Assessment Program: Findings and Implications" by Koretz, Stecher, Klein, and McCaffery (1994).

agenda for the last 20 years (1975, 1989a, 1989b, 1994). Others have since added their support (Cherryholmes, 1988; Berlak, 1992; Moss, 1992; Shepard, 1993). In a challenge to the *Standards* (AERA, APA, & NCME, 1985), Berlak and his co-authors (1992) have proposed an alternative paradigm to take account of the many contextual influences that impact upon assessment design and interpretation.

Many are looking to the revision of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1985) for guidance in a wide range of contexts. Linn, writing in the December 1994 edition of *Educational Researcher*, identifies three additional factors that will impact upon the design of the "new" *Standards*. These are (1) the political agenda of the U.S. federal government, (2) the creation of U.S. national standards, and (3) the expanded reliance upon performance-based assessment. These *Standards* will be used extensively in Canada where the parallel pressures are (1) the expanded interest of the provincial governments in setting and monitoring standards[2], and (2) greater use of performance-based assessment. Opportunities exist for Canadian measurement professionals to become involved in the revisions of the *Standards* at meetings or by written submission as took place with the *Program Evaluation Standards* (Sage, 1994).

The emerging literature in science education standards is linked with the revisions in measurement standards. Many changes in standards are proposed. Some have moved beyond a statement to proceed and are already at the response stage, for example the *Draft of the National Science Education Standards* (National Research Council, 1994). Others,

---

2 The B.C. Ministry of Education identifies meeting "standards for achievement in curriculum" as part one of its current (September, 1994) goals of education. Similarly the Ontario Royal Commission on Learning recommends that clearly written standards be developed in science (January, 1995).

such as the provincial curriculum standards in B.C. are in the process of preparation for the consultation process. The influence of the new American national education standards will inevitably permeate north of the border. There are a wide range of emphases: teaching, professional development, assessment, content, program and system. One of the key areas of emphasis in the new science curriculum documents is the enabling of students to participate in and experience scientific inquiry. This is an area where performance tasks offer significant potential for direct assessment in the form of investigations. The positive experiences of the students and the teacher/administrators to the open ended investigations used in the Student Performance Component of the 1991 B.C. Science Assessment attest to the value of such an approach to assessment.

The expanding future for hands-on performance assessment tasks in science offers a broad range of research opportunities. Three areas follow directly from this study. The first relates to the use of the validation framework to evaluate claims and the consequential effects of the use of performance assessments in a context other than that presented here in Chapter 4. The second area is to use the model template illustrated in Chapter 5 to examine content representation, in conjunction with generalizability studies, to predict task-sampling variability. This will aid in the design of assessments that are sensitive to such effects. Research should also be directed towards a comparative analysis of holistic and analytical scoring systems, with the explicit intent of providing some guidelines to suggest under what circumstances they can be synthesized and when a particular approach might illuminate student performance with greater clarity.

REFERENCES

American Association for the Advancement of Science (1965). *Science a process approach.* Xerox.

American Psychological Association (1954). *Technical recommendations for psychological tests and diagnostic techniques.* Washington, DC: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Anastasi, A. (1988). *Psychological testing.* (6th ed.) New York, NY: Macmillan.

Anderson, R. (1990). *California: The state of assessment.* Sacramento: California State Department of Education.

Archbald, D. A., & Newmann, F. M. (1988). *Beyond standardized tests: Assessing authentic academic achievement in the secondary school.* Reston, VA: National Association of Secondary School Principals.

Archenhold, W. F., Bell, J., Donnelly, J., Johnson, S., & Welford, G. (1988). *Science at age 15: a review of A.P.U. findings 1980-84.* London: HMSO.

Baron, J. B. (1991). Performance assessment: blurring the edges among assessment, curriculum and instruction. In G. Kulm, & S. Malcolm (Eds.), *Assessment in the service of reform.* (pp. 247-266). Washington D.C.: American Association for the Advancement of Science.

Bartley, A. W. (1991). *Student performance tasks: Administration manual.* Vancouver, BC: University of British Columbia.

Bartley, A. W., Carlisle, R. W., & Erickson, G. (1993). *Science program assessment through performance assessment tasks: A guide for school districts.* Victoria, BC: Ministry of Education and Ministry Responsible for Multiculturalism and Human Rights.

Bateson, D. J., Anderson, J., Brigden, S., Day, E., Deeter, B., Eberlé, C., Gurney, B., & McConnell, V. (1992). *British Columbia Assessment of Science 1991 Technical Report I: Classical Component.* Victoria, BC: Ministry of Education and Ministry Responsible for Multiculturalism and Human Rights.

Bateson, D. J., Anderson, J. O., Dale, T., McConnell, V., & Rutherford, C. (1986). *Science assessment 1986, general report.* Victoria: Ministry of Education.

Bateson, D. J., Erickson, G., Gaskell, P. J., & Wideen, M. (1992). *British Columbia assessment of science: Provincial report 1991.* Victoria, BC: Ministry of Education and Ministry Responsible for Multiculturalism and Human Rights.

Bateson, D. J., & Parsons, S. (1989). Sex-related differences in science achievement: a possible testing artifact. *International Journal of Science Education,* 11 (4), 371-385.

Bathory, Z. (1985). The CTD science practical survey. *Studies in Educational Evaluation,* (9), 165-174.

Baxter, G. P., Shavelson, R. J., Goldman, S. R., & Pine, J. (1992). Evaluation of procedure-based scoring for hands-on assessment. *Journal of Educational Measurement,* 29 (1), 1-17.

Baxter, G. P., Shavelson, R. J., Herman, S. J., Brown, K. A., & Valadez, J. R. (1993). Mathematics performance assessment: Technical quality and diverse student impact. *Journal for Research in Mathematics Education,* 24 (3), 190-216.

Bereiter, C. (1985). Towards a solution of the learning paradox. *Review of Educational Research,* 55 (2), 201-206.

Berlak, H. (1992). The need for a new science of assessment. H. Berlak, F. M. Newmann, E. Adams, D. A. Archbald, T. Burgess, J. Raven, & T. Romberg (eds.), *Toward a new science of educational testing and measurement.* (pp. 1-21). Albany, NY: State University of New York Press.

Berlak, H., Newmann, F. M., Adams, E., Archbald, D. A., Burgess, T., Raven, J., & Romberg, T. (1992). *Toward a new science of educational testing and measurement.* Albany, NY: State University of New York.

Black, P. (1986). Integrated or coordinated science? *School Science Review,* 67 (241), 669-681.

Bloom, B. (1956). *Taxonomy of educational objectives.* New York, NY: David Mckay.

Blumberg, F., Epstein, M., MacDonald Walter, & Mullis Ina. (1986). *A pilot study of higher order thinking skills assessment techniques in science and mathematics – final report.* Princeton: National Assessment of Educational Progress.

Brennan, R. L. (1983). *Elements of generalizability theory.* Iowa City, IA: American College Testing Program.

British Columbia Department of Education (1975). *Assessment planning: B.C. assessment program.* Victoria B.C.: Author.

Bryce, T. J. K., McCall, J., MacGregor, J., Robertson, I. J., & Weston, R. A. (1984). *Techniques for the assessment of practical skills in foundation science: report of the project (1980-1983).* Glasgow: Jordanhill College of Education.

Bryce, T. J., & Robertson, I. J. (1985). What can they do? A review of practical assessment in Science. *Studies in Science Education,* 12 , 1-24.

Bryce, T. J., & Robertson, I. J. (1986). Practical science assessment. Will it work in England and Wales? *The Times Educational Supplement,* (18.04.86), 63.

Burger, S. E., & Burger, D. L. (1994). Determining the validity of performance-based assessment. *Educational Measurement: Issues and Practice,* 13 (1), 9-15.

California Department of Education (1990). *California science framework.* Sacramento, CA: Author.

Candell, G. L., & Ercikan, K. (1992). *Assessing the reliability of the Maryland School Performance Assessment Program using generalizability theory.* : Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco.

Carnegie Commission on Science Technology and Government. (1991). *In the national interest: The federal government in the reform of K-12 math and science education.* New York, NY: Author.

Carnegie Forum on Education and the Economy. (1986). *A nation prepared: Teachers for the 21st century.* New York, NY: Carnegie.

Champagne, A. B., Lovitts, B. E., & Callinger, B. J. (1990). *This year in school science 1990: Assessment in the service of instruction.* Washington D.C: American Association for the Advancement of Science.

Cherryholmes, C. (1989). *Power and criticisms; poststructural investigations in education.* New York, NY: Teachers College Press.

Churchman, C. W. (1971). *The design of inquiring systems: Basic concepts of systems and organizations.* New York, NY: Basic Books.

Cole, N. S. (1991). The impact of science assessment on classroom practice. In G. Kulm, & S. Malcolm (Eds.), *Science assessment in the service of reform.* (pp. 97-105). Washington D.C.: American Association for the Advancement of Science.

Cole, N. S., & Moss, P. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement.* (3rd) (pp. 201-219). New York, NY: Macmillan.

Comfort, K. B. (1990). *New Directions in Science Assessment.* Sacramento: California Department of Education.

Comfort, K. B. (1994). *A sampler of science assessment: Elementary.* Sacramento: California Department of Education.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York, NY: Holt, Reinhart and Winston.

Cronbach, L. J. (1970). *Essentials of psychological testing.* (3rd ed.) New York, NY: Harper and Row.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement.* (2nd ed.) (pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer, & H. I. Braun (Eds.), *Test validity.* (pp. 3-17). Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1989). Construct validity after thirty years. In R. L. Linn (Ed.), *Intelligence: measurement theory and public policy.* (pp. 147-171). Urbana: University of Illinois Press.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements.* New York, NY: Wiley.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin,* 52 , 281-302.

Dabney, V. M. *How can we insure accurate and reliable data from authentic assessments, or can we?* Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA: 1993.

Department of Education and Science and the Welsh Office. (1988). *Science in the national curriculum.* London: HMSO.

Department of Education and Science and the Welsh Office. (1991). *Science in the national curriculum (1991).* London: HMSO.

Donnelly, J. F., & Gott, R. (1985). An assessment-led approach to processes in the science curriculum. *European Journal of Science Education,* 7 (3), 237-251.

Doran, R. L., & Tamir, P. (1992). Results of practical skills testing. *Studies in Educational Evaluation,* 18 (1), 393-406.

Driver, R., Gott, R., Johnson, S., Worsley, C., & Wylie, F. (1982). *Science in schools. Age 15: Report no. 1.* London: HMSO.

Driver, R., Guesne, E., & Tiberghien, A. (1985). (Eds.), *Children's ideas in science.* Milton Keynes: Open University Press.

Erickson, G. (1990). *Report to ministry of education on the assessment of students' practical work in science.* Vancouver: The University of British Columbia.

Erickson, G., Bartley, A. W., Blake, L., Carlisle, R. W., Meyer, K., & Stavy, R. (1992). *British Columbia assessment of science 1991 technical report II: Student performance component.* Victoria, BC: Ministry of Education and Ministry Responsible for Multiculturalism and Human Rights.

Erickson, G., Carlisle, R. W., & Bartley, A. W. (1992). Performance assessment component. In D. J. Bateson, G. Erickson, P. J. Gaskell, & M. Wardeen (Eds.), *British Columbia assessment of science: Provincial report 1991.* (pp. 25-40). Victoria, BC: Ministry of Education and Ministry Responsible for Multiculturalism and Human Rights.

Erickson, G., & Farkas, S. (1991). Prior experiences and gender differences in science achievement. *Alberta Journal of Educational Research,* XXXVII (3), 225-239.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement.* (3rd) (pp. 105-146). New York, NY: Macmillan.

Feyerabend, P. (1975). *Against Method.* London: New Left Books.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher,* 18 (9), 27-32.

Gagné, R. M. (1965). *The psychological basis of science - a process approach.* (pp. 65 - 68). Washington D.C.: American Association for the Advancement of Science.

Gaskell, P. J., Fleming, R., Fountain, R., & Ojelel, A. (1992). *British Columbia Assessment of Science 1991 Technical Report III: Socioscientific Component.* Victoria, BC: Ministry of Education and Ministry Responsible for Multiculturalism and Human Rights.

Glaser, R. (In press). Testing and assessment: O Tempora! O Mores! *Educational Researcher,*

Gott, R., & Murphy, P. (1987). *Assessing investigations in science, ages 13 and 15.* London: HMSO.

Greeno, J. G. (1989). A perspective on thinking. *American Psychologist,* 39 , 193-202.

Hanna, G. S. (1993). *Better teaching through better measurement.* Fort Worth, TX: Harcourt Brace Jovanovich.

Hardy, R. (1992). *Options for scoring performance assessment tasks.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco.

Harlen, W., Palacio, D., & Russell, T. (1984). *The APU assessment framework for science at age 11.* London: HMSO.

Hein, G. E. (1990). Conclusion. In G. E. Hein (editor.), *The assessment of hands-on elementary programs.* (pp. 264-279). Grand Forks, ND: Center for Teaching and Learning, University of North Dakota.

Herman, J. L. (1992). What research tells us about good assessment. *Educational Leadership,* 49 (8), 74-78.

Hieronymous, A. N., & Hoover, H. D. (1987). *Iowa Test of Basic Skills: Writing supplement teacher's guide.* Chicago, IL: Riverside.

Hobbs, E. D., Boldt, W. B., Erickson, G. L., Quelch, T. P., & Sieben, G. A. (1980). *British Columbia science assessment 1978: General report, volume I: Procedures, student Test, conclusions and recommendations.* Victoria: Ministry of Education.

Hodson, D. (1986). The nature of scientific observation. *School Science Review,* 68 (242), 17-29.

Howe, K. R. (1985). Two dogmas of educational research. *Educational Researcher,* 14 (8), 10-18.

Johnson, S. (1987). Gender differences in science: parallels in interest, experience and performance. *International Journal of Science Education,* 9 (3), 467-481.

Johnson, S. (1989). *National Assessment: The APU science approach.* London: HMSO.

Kahle, J. B. (1988). Gender and science education II. In P. Fensham (Ed.), *Development and dilemmas in science education.* (pp. 249-265). Lewes, Sussex: Falmer Press.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin,* 112 , 527-535.

Kanis, I. B., Doran, R. L., & Jacobson, W. J. (1990). *Assessing laboratory process skills at the elementary and middle/high levels.* New York: The Second International Science Study, Teachers College, Columbia University. Available through NSTA, Washington DC 20009.

Kaplan, A. (1964). *The conduct of inquiry: Methodology for behavioral science.* San Francisco, CA: Chandler and Sharp.

Kempa, R. (1986). *Assessment in science.* Cambridge: Cambridge University Press.

Klinger. D. (1994) *Science performance assessment scoring guide.* Langley, B.C.: Langley School District.

Klopfer, L. E. (1971). Evaluation of learning in science. In B. S. Bloom, J. T. Hastings, & G. F. Madaus (Eds.), *Handbook on formative and summative evaluation of student learning.* New York, NY: McGraw-Hill.

Koretz, D., Stecher, B., Klein, S., McCaffery, D., & Deibert, E. (1992). *Can portfolios assess student performance and influence instruction? The 1991-92 Vermont experience.* RAND Institute on Education and Training.

Koretz, D., Stecher, B., Klein, S., & McCaffery, D. (1994). The Vermont portfolio assessment program: findings and implications. *Educational Measurement: Issues and Practice,* 13 (3), 5-16.

Lane, S. (1994). Book review: Toward a new science of educational testing and measurement. *Educational Measurement: Issues and Practice,* 13 (1), 40-43.

Lane, S., Parke, C., & Moskal, B. (1992). *Principles for developing performance assessments.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco.

Linn, M., & Hyde, J. S. (1989). Gender, mathematics, and science. *Educational Researcher,* November , 17-27.

Linn, R. L. (Ed.). (1989). *Educational measurement.* (3rd ed.) New York, NY: Macmillan.

Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher,* 23 (9), 4-14.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex performance-based assessment: expectations and validation criteria. *Educational Researcher,* 20 (8), 15-21.

Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice,* 13 (1), 5-15.

Lovitts, B. E., & Champagne, A. B. (1990). Assessment and instruction: two sides of the same coin. In A. B. Champagne, B. E. Lovitts, & B. J. Callinger (Eds.), *This year in school science 1990: Assessment in the service of instruction.* (pp. 1-13). Washington D.C: American Association for the Advancement of Science.

Loyd, B. H. (1994). Book review: Toward a new science of educational testing and measurement. *Journal of Educational Measurement,* 31 (1), 83-87.

Magone, M., Cai, J., Silver, E. A., & Wang, N. (1992). *Validity evidence for cognitive complexity of performance assessments: An analysis of selected QUASAR tasks.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco.

Manitoba Curriculum Development and Assessment Branch (1988). *Manitoba science assessment 1986: Final report.* Winnipeg: Author.

McCombs, B. L. (1991). The definition and measurement of primary motivational processes. In M. C. Wittrock, & E. L. Baker (Eds.), *Testing and cognition.* Englewood Cliffs, NJ: Prentice Hall.

Mehrens, W. A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice,* 11 (1), 3-9.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist,* 30 , 955-966.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher,* 23 (2), 13-23.

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

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

Meyer, C. A. (1992). What's the difference between authentic and performance assessment? *Educational Leadership,* 49 (8), 39-40.

Meyer, K. (1991). Children as experimenters: Elementary students' actions in an experimental context with magnets. Unpublished doctoral dissertation, University of British Columbia, Vancouver.

Millar, R. (1989). What is 'scientific method' and can it be taught? In J. J. Wellington (Ed.), *Skills and processes in science education: A critical appraisal.* (pp. 47-62). London: Routledge.

Millar, R. (1991). A means to an end: the role of processes in science education. In B. Woolnough (Ed.), *Practical Science.* (pp. 43-52). Milton Keynes: Open University Press.

Millar, R., & Driver, R. (1987). Beyond processes. *Studies in Science Education,* (14), 33 -62.

Miller-Jones, D. (1989). Culture and testing. *American Psychologist,* 44 , 360-366.

Ministry of Education (1981). *Elementary science curriculum guide: Grades 1-7.* Victoria, B.C.: Author.

Ministry of Education (1985). *Junior secondary science: curriculum guide and resource book.* Victoria, B.C.: Author.

Ministry of Education (1990). *Year 2000: A framework for learning.* Victoria, B.C.: Author.

Ministry of Education (1991). *Primary Program: foundation document.* Victoria, B.C.: Author.

Ministry of Education (1994). *The kindergarten to grade 12 education plan.* Victoria, BC: Province of British Columbia.

Ministry of Education and Ministry Responsible for Multiculturalism and Human Rights (1992a). *Ministry response the 1991 British Columbia Assessment of Science.* Victoria, BC: Ministry of Education and Ministry Responsible for Multiculturalism and Human Rights.

Ministry of Education and Ministry Responsible for Multiculturalism and Human Rights (1992b). *Curriculum and assessment framework: science.* Victoria, BC: Ministry of Education and Ministry Responsible for Multiculturalism and Human Rights.

Mitroff, I. I., & Sagasti, F. (1973). Epistemology as general systems theory: An approach to the design of complex decision-making experiments. *Philosophy of Social Sciences,* (3), 117-134.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research,* 62 (3), 229-258.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher,* 23 (2), 5-12.

Mullis, I. V. S. (1980). *Using the primary trait system for evaluating writing.* Denver, CO: Education Commission of the States.

Murphy, P. (1989). Across-category performance issues. In B. Schofield, P. Black, J. Bell, S. Johnson, P. Murphy, A. Qualter, & T. Russell, *Science at age 13: a review of APU findings 1980-84.* (pp. 149-157). London: HMSO.

Murphy, P. (1990). What has been learned about assessment from the work of the APU science project? In G. E. Hein (editor.), *The assessment of hands-on elementary programs.* (pp. 148-179) Grand Forks, ND: Center for Teaching and Learning, University of North Dakota.

Murphy, P., & Gott, R. (1984). *The assessment framework for science at ages 13 and 15.* (Science report for teachers 3) London: Department of Education and Science.

National Center on Education and the Economy. (1990). *America's choice: High skills or low wages!* Rochester, NY: National Center on Education and the Economy.

National Commission on Excellence in Education. (1983). *A nation at risk.* Washington, DC: U.S. Department of Education.

National Committee on Science Education Standards and Assessment. (1993). *National science education standards: An enhanced sampler.* Washington, DC: National Research Council.

National Governors' Association. (1986). *A time for results: The governors' report on education.* Washington, DC: National Governors' Association.

National Research Council. (1994). *National science education standards: Draft for review.* Washington, DC: National Academy Press.

New York State Department of Education (1989). *Teachers' guide to administration of grade 4 performance tasks.* Albany: Author.

Newmann, F. M. (1991). Linking restructuring to authentic student achievement. *Phi Delta Kappan,* (February), 458-463.

Newmann, F. M., & Archbald, D. A. (1992). The nature of authentic academic achievement. In H. Berlak, F. M. Newmann, E. Adams, D. A. Archbald, T. Burgess, J. Raven, & T. Romberg (Eds.), *Towards a new science of educational testing and assessment.* Albany, NY: State University of New York Press.

Osborne, R., & Freybourg, P. (1985). *Learning in science - The implications of children's science.* Auckland: Heinemann.

Pine, J. (1990). Validity of science assessments. In G. E. Hein (editor.), *The assessment of hands-on elementary programs.* (pp. 83-94). Grand Forks, ND: Center for Teaching and Learning, University of North Dakota.

Popham, W. J. (1987). The merits of measurement-driven instruction. *Phi Delta Kappan,* 68 , 679-682.

Postlethwaite, T. N., & Wiley, D. E. (Eds). (1992). *Science achievement in twenty-three countries.* Oxford: Pergamon Press.

Raven, J. (1992). A model of competence, motivation, and behavior, and a paradigm for assessment. In H. Berlak, F. M. Newmann, E. Adams, D. A. Archbald, T. Burgess, J. Raven, & T. Romberg, *Toward a new science of educational testing and measurement.* (pp. 85-116). Albany, NY: State University of New York Press.

Raymond, M. R., & Houston, W. M. (1990, April). *Detecting and correcting for rater effects in performance assessment.* Paper presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, Boston, MA.

Resnick, L. B. (1983). Mathematics and science learning: A new conception. *Science,* (220), 477-478.

Roberts, D. (1988). What counts as science education? In P. J. Fensham (Ed.), *Development and dilemmas in science education.* (pp. 27-54). Lewes, England: Falmer Press.

Robertson, I. J. (1987). Girls and boys and practical science. *International Journal of Science Education,* 9 (5), 505-518.

Robertson, I. J. (1994). *Making inferences and evaluating evidence in practical investigations.* Paper presented at the Annual Meeting of National Association for Research in Science Teaching, Anaheim, CA.

Robitaille, D. F., Schmidt, W. H., Raizen, S., McKnight, C., Britton, E., & Nicol, C. C. (1993). *TIMSS Monograph No. 1: Curriculum frameworks for mathematics and science.* Vancouver, BC: Pacific Educational Press.

Romberg, T., & Zarinnia, A. (1987). Consequences of the new world view to assessment of students' knowledge in mathematics. In T. Romberg, & D. Stewart (Eds.), *The monitoring of school mathematics: Background papers. Vol. 2. Implications from psychology: Outcomes of instruction.* Wisconsin: University of Wisconsin-Madison.

Rothman, R. (1990 a). New tests based on performance raise questions - assessment methods said like star wars. *Education Week,* (Volume X, Number 2), 1,10 & 12.

Rothman, R. (1990 b). U.S. opts out of international performance assessment. *Education Week,* (Volume X, Number 2), 10.

Ruiz-Primo, M. A., Baxter, G. P., & Shavelson, R. J. (1993). On the stability of performance assessments. *Journal of Educational Measurement,* 30 (1), 41-53.

Russell, T., Black, P., Harlen, W., Johnson, S., & Palacio, D. (1988). *Science at age 11: a review of APU findings 1980-84.* London: HMSO.

Sage Publications Inc. (1994). *The program evaluation standards.* Thousand Oaks, CA: Sage Publications.

Schofield, B., Black, P., Bell, J., Johnson, S., Murphy, P., Qualter, A., & Russell, T. (1989). *Science at age 13: a review of APU findings 1980-84.* London: HMSO.

Semple, B. (1992). *Performance assessment: an international experiment.* Princeton: International Assessment of Educational Progress.

Shavelson, R. J., & Baxter, G. P. (1992). What we've learned about assessing hands-on science. *Educational Leadership,* 49 (8), 20-25.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement,* 30 (3), 215-232.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: political rhetoric and measurement reality. *Educational Researcher,* 21 (4), 22-27.

Shavelson, R. J., Baxter, G. P., Pine, J., Yur , J., Goldman, S. R., & Smith, B. (1991). Alternative technologies for large-scale science assessment: Instruments of educational reform. *School Effectiveness and School Improvement,* 2 (2), 97-114.

Shavelson, R. J., Gao, X., & Baxter, G. P. (1994). On the content validity of performance assessments: Centrality of domain specification. An invited paper at the *First European Electronic Conference on Assessment and Evaluation.* : Conference on the Internet by The European Association for Research on Learning and Instruction (EARLI).

Shavelson, R. J., & Webb, N. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage Publications.

Shepard, L. (1991). Psychometricians' beliefs about learning. *Educational Researcher,* 20 (7), 2-16.

Shepard, L. (1993). Evaluating test validity. *Review of Research in Education,* (19), 405-450.

Slater, T. F., & Ryan, J. M. (1993). Laboratory performance assessment. *The Physics Teacher,* 31 , 306-308.

Snow, R. E. (1974). Representative and quasi-representative designs for research on teaching. *Review of Educational Research,* 44 , 265-291.

Sullivan, B. M. (1987-88). *A legacy for learners: the report of the royal commission on education.* Victoria: Ministry of Education.

Swain, J. R. L. (1974). Practical objectives - A review. *Education in Chemistry,* 11 (5), 152-156.

Tamir, P., & Doran, R. L. (1992a). Scoring guidelines. *Studies in Educational Evaluation,* 18 (1), 355-363.

Tamir, P., & Doran, R. L. (1992b). Conclusions and discussion of findings related to practical skills testing in science. *Studies in Educational Evaluation,* 18 , 393-406.

Tamir, P., Doran, R. L., & Chye, Y. O. (1992). Practical skills testing in science. *Studies in Educational Evaluation,* 18 (1), 263-275.

Tamir, P., Doran, R. L., Kojima, S., & Bathory, Z. (1992). Procedures used in practical skills testing in science. *Studies in Educational Evaluation,* 18 (1), 277-290.

Tamir, P., & Lunetta, V. N. (1978). An analysis of laboratory activities in BSCS yellow version. *American Biology Teacher,* (40), 426-428.

Taylor, H., Hunt, R., Sheppy, J., & Stronck, D. (1982). *Science assessment 1982, general report.* Victoria: Ministry of Education.

Traub, R. E., & Rowley, G. L. (1991). Understanding reliability. *Educational Measurement: Issues and Practice,* 10 (1), 37-45.

Tucker, M. S. (1991). Why assessment is now issue number one. In G. Kulm, & S. Malcolm (Eds.), *Science assessment in the service of reform.* (pp. 3-15). Washington D.C.: American Association for the Advancement of Science.

United States Department of Education (1990). *National goals for education.* Washington, DC: Author.

Welch, C. J. (1993). *Issues in developing and scoring performance assessments.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta, GA.

Welford, G., Harlen, W., & Schofield, B. (1985). *Practical testing at ages 11, 13 and 15.* London: HMSO.

Whittaker, R. J. (1974). The assessment of practical work. In Macintosh H. G. (Ed.), *Techniques and problems in assessment.* London: Arnold.

Wideen, M., Mackinnon, A., O'Shea, T., Wild, R., Shapson, S., Day, E., Pye, I., Moon, B., Cusack, S., Chin, P., & Pye, K. (1992). *British Columbia Assessment of Science 1991 Technical Report IV: Context for Science Component.* Victoria, BC: Ministry of Education and Ministry Responsible for Multiculturalism and Human Rights.

Wienekamp, H., Jansen, W., Fickenferichs, H., & Peper, R. (1987). Does the unconscious behaviour of teachers cause chemistry lessons to be unpopular with girls? *International Journal of Science Education,* 9 (3), 281-286.

Wilson, G. (1986). *CHASSIS: Cheshire achievement of scientific skills in schools.* Chester, England: Cheshire County Council.

Wittrock, M. C. (1991). Testing and recent research in cognition. In M. C. Wittrock, & E. L. Baker (Eds.), *Testing and cognition.* Englewood Cliffs, NJ: Prentice Hall.

Woolnough, B. (Ed.). (1991). *Practical Science.* Milton Keynes: Open University Press.

Woolnough, B. (1989). Towards a holistic view of science education (or the whole is greater than the sum of its parts, and different. In J. J. Wellington (Ed.), *Skills and processes in science education: A critical appraisal.* London: Routledge.

Woolnough, B., & Allsop, T. (1985). *Practical work in science.* Cambridge: Cambridge University Press.