

**Investigating the Statistical and Cognitive Dimensions in
Large-Scale Science Assessments: Causal and Categorical Reasoning in Science**

Jacqueline P. Leighton

Rebecca J. Gokiert*

Ying Cui

Center for Research in Applied Measurement and Evaluation (CRAME)

University of Alberta, CANADA

***Presenting Author. Paper presented at the Annual Meeting of the American Educational Research Association (AERA), Montreal, Quebec, Canada (April 2005).**

Abstract

In a special issue of the journal *Educational Assessment* titled ‘A Multidimensional Approach to Achievement Validation,’ a series of studies were presented that focused on extending and elaborating the work of Richard E. Snow (e.g., Roeser et al., 2002; Shavelson et al., 2002). The studies in this special issue are compelling in that they demonstrate the multidimensional nature of science achievement across a variety of large-scale achievement tests (NELS, TIMMS, and NAEP) using factor analysis (see also Nussbaum, Hamilton, Snow, 1997). The current study was conducted in two steps. The first step was to (a) identify the dimensional structure of a new large-scale science assessment using a non-parametric technique, such as DIMTEST (Stout, Douglas, Junker, & Roussos, 1993), in addition to exploratory factor analysis with a variety of factor analytic decision rules (Zwick & Velicer, 1986), and the second step was to (b) identify and describe the cognitive psychological attributes of the dimensions found to underlie this new large-scale science assessment. This second objective is focused on providing increasing support for the view of dynamic performance in science achievement. The results of the present study indicate that science assessments involve at least two substantive dimensions to which students react—causal reasoning and categorical reasoning—described in the scientific reasoning literature (Kuhn & Deane, 2004).

Investigating the Statistical and Cognitive Dimensions in Large-Scale Science Assessments: Causal and Categorical Reasoning in Science

If you think back to the last time you went grocery shopping and tried to calculate the lowest price for a box of cereal—the 450g box for 3.99 or the 1Kg box for 6.99—you can appreciate the give-and-take effort required to settle on an answer. Now imagine the substantial give-and-take of solving a stoichiometry or physics question. What is the question asking the problem solver to do? What are the important variables? What kinds of knowledge or rules need to be used? Is there time to solve the problem? From this example, it is possible to appreciate that problem solving is not defined by the task alone but also by the problem solver's interpretation of the task; in other words, problem solving is an interactive process between the problem solver and his or her environment (Giroto, 2004; Leighton & Sternberg, 2003). The interactive nature of problem solving has led to the description of academic achievement as *dynamic* (Dixon, 1992; Fischer & Pipp, 1984; Gardner, 1993; Snow & Lohman, 1989; Roeser et al., 2002; Shavelson et al., 2002), and educational assessments and/or achievement tests as *multidimensional* (Nussbaum, Hamilton, & Snow, 1997; Roeser et al., 2002; Shavelson et al., 2002).

In a special issue of the journal *Educational Assessment* titled 'A Multidimensional Approach to Achievement Validation,' a series of papers were presented that focused on extending and elaborating the work of the late Richard E. Snow (e.g., Roeser et al., 2002; Shavelson et al., 2002). This special issue is a reminder of Snow's legacy in educational measurement: The view of test performance as a dynamic interaction between a student's aptitudes and a test's contextual features or characteristics, thus leading to a multidimensional latent test structure. The papers in this special issue are compelling for at least two reasons. First, using a variety of statistical tools, the papers (Ayala, Shavelson, Yin, & Schultz, 2002; Kuppermintz, 2002; Lau & Roeser, 2002; Haydel & Roeser, 2002; Roeser et al., 2002; Shavelson et al., 2002) converge in their conclusion that academic performance across a variety of large-scale achievement tests (TIMMS and NAEP) is *dynamic* and achievement tests are *multidimensional*. These studies support and build on a factor analytic study, conducted by Snow and his colleagues (i.e., Hamilton, Nussbaum, & Snow, 1997; Nussbaum, Hamilton, & Snow, 1997), which demonstrated that three latent traits could be used to describe performance on the science section of the National Education Longitudinal Study of 1988 (NELS:88). Second, these papers emphasize the important link between cognitive, conative, and affective attributes with student test performance.

The term *dynamic* has a theoretically laden meaning. Traditionally, the dynamic systems approach has been used to describe phenomena in biology and physics (Pfeifer & Scheier, 1999; Thagard, 1996). In particular, to describe how a biological or physical system changes over time to a set of environmental variables. More recently, cognitive scientists have adopted the dynamic system approach to explain human cognition and its unpredictable nature across situations (Holland, 1998; Pfeifer & Scheier, 1999; Thagard, 1996; van Geert, 2003). Paul van Geert (2003, p. 198) maintains that the core belief of the dynamic view (as applied to human cognition) is "that intelligence is assembled, each time anew, in the ongoing, time-dependent interaction between properties of the person (abilities or effectivities) and properties of the physical, symbolic, and social environment."

Dynamic system theory functions as a meta-theory in educational testing, indicating the importance of considering the student attributes and environmental variables that might interact to produce behaviour. Although there is no single theory that can be called a dynamic theory of science achievement at this time, dynamic system theory specifies a class of theories that involve certain priorities when describing student performance. For example, a theory of science achievement that merits the descriptor of dynamic should describe the test variables that students find meaningful and therefore differentially react to. Moreover, the test variables identified should be described within a larger cognitive theory of reasoning and problem solving so as to understand the credence of these variables. Generating a theory that merits the descriptor dynamic is not an

easy task given the multitude of variables that can impinge on students' performance at any one time. However, identifying these variables is a necessary task according to the dynamic systems approach. Identifying and describing why students might respond differentially to specific variables is required to build dynamic models, which are then subjected to empirical testing. A leading cognitive scientist, John H. Holland (1998) describes this process:

To build a dynamic model we have to select a level of detail that is useful, and then we have to capture laws of change at that level of detail. There are potential conflicts. It may be quite difficult to construct a detailed model that is "faithful" to the system being modeled. Weather prediction models provide instructive examples. Predicting that the atmospheric temperature will be less than the boiling point of water may be reassuring, but it is not much of a weather prediction. *We want more detail, but then we have to deal with laws of change that involve fronts, jet streams, and the like.* (p. 46, italics added)

To describe performance as dynamic entails a commitment to describing the specific environmental variables that affect behavior in student performance. Furthermore, dynamic models of performance must be created and ultimately evaluated at a level of detail specific enough to indicate which variable is responsible for which student action. Identifying the potential causes of behavior is what lies at the core of dynamic descriptions. This requirement of dynamic models suggests that when we make the claim "student test performance is dynamic," we must be ready to be specific about the details of the environmental variables responsible for the performance. In short, this requirement strongly suggests that we have to do a better job describing the variables that may be affecting behavior in students.

The term *dynamic* has come to be associated with multidimensional assessments; that is, assessments that measure multiple traits or abilities in students (e.g., Roeser et al., 2002; Shavelson et al., 2002). The terms dynamic and multidimensional can be viewed as complementary because one would expect that assessments involving multiple dimensions should elicit distinct behaviors from students; a student responds differently depending on the dimension encountered, thus generating an interaction between student and test. When this is the case, we must attempt to identify and describe the underlying test dimensions with enough detail (i.e., the environmental variables) so as to hypothesize which relevant variables are involved in the interaction between student and test.

If one accepts the sensibility of viewing test performance as dynamic, then it would appear that at a minimum we must be able to show that a test is multidimensional; in other words, the test's dimensions are distinct enough to produce distinct behaviors from students. The aim of the present study is to identify the dimensional structure of a large-scale science assessment called the School Achievement Indicators Program (SAIP) Science Assessment. The first objective of the current study was to use DIMTEST (Ackerman, Gierl, & Walker, 2003; Stout, Douglas, Junker, & Roussos, 1993), in addition to exploratory factor analysis (Zwick & Velicer, 1986), to determine the dimensional structure of the SAIP Science 1999 Assessment. The second objective was to use the exploratory results to identify a cognitive model of test performance, and to test the model using confirmatory factor analysis. Identifying a cognitive model provides information about the test variables that might be eliciting distinct behaviors in students. The paper is divided into six sections: first, we provide a rationale for viewing performance as dynamic and science assessments as multidimensional; second, we present our rationale for the exploratory analysis of the SAIP Science Assessment; third, we describe in detail the SAIP Science Assessment used in the present study; fourth, we present the results from a series of exploratory analyses aimed at uncovering the underlying latent structure of SAIP; fifth, we interpret substantively the latent structure of the assessment and compare our results with results from past studies; sixth, we test the set of cognitive models hypothesized to underlie the SAIP Assessment. We summarize the paper by

discussing the usefulness of dimensionality studies in shifting the way in which testing is currently conceptualized.

Dynamic Performance and Multidimensional Science Tests

Given the recent psychological literature on problem solving, and specifically the literature on expert-novice thinking, one would be prudent in concluding that problem solving is dynamic; cognitive psychologists frequently acknowledge the empirical evidence that demonstrates the interaction between person and task (Charness & Schultetus, 1999; Girotto, 2004; Kuhn, 2001; Leighton & Sternberg, 2003; Thagard, 1996; van Geert, 2003). However, it is not sufficient to say that performance is dynamic without identifying on a given occasion the specific task variables contributing to the performance of interest. In other words, for a test or class of tests to be viewed as eliciting dynamic performance, an argument must be created for substantiating the label dynamic performance. According to Kane's (2001, pp. 329-330) explication of the argument-based approach to validity, "the interpretive argument will generally contain a number of inferences and assumptions (as all arguments do), and the studies to be included in the validation effort are those studies that are most relevant to the inferences and assumptions in the specific interpretive argument under consideration."

The argument for suggesting that a given test elicits dynamic performance will probably contain more inferences and assumptions about the test's cognitive complexity and fewer inferences and assumptions about the test's content standards (Cronbach, 1988; Kane, 2001; Lane, 2004). According to Lane (2004), content standards should not be used as a surrogate for a more thorough, cognitive description of the factors found to underlie a test. Content standards often fail to represent the cognitive complexity (Lane, 2004) of test items. If content standards are used alone, without appeal to psychological or cognitive processes, then the interpretation of factors will not represent the dynamic interplay between student and test. These content-based descriptions reflect test developers' ideas about what the test is measuring and may not necessarily reflect students' interpretations (Kane, 2001, p. 320; Leighton, 2004; Norris, Leighton & Phillips, 2004). Test items can be grouped according to the content and skills indicated in the test specifications, but these content and skills may not have any psychological reality for students taking the test. The path to creating dynamic models of test performance begins by focusing on the cognitive psychological characteristics of underlying test dimensions produced or elicited by students. Furthermore, by focusing on the cognitive psychological characteristics of the underlying test dimensions, Kane's (2001) strong program of construct validity can be pursued. The cognitive characteristics are used to build a dynamic model of science achievement rooted in psychological theory that is open to empirical scrutiny:

The strong program outlined by Cronbach and Meehl (1955) has a narrower focus but it has teeth. One is to lay out theoretical assumptions and conclusions and then subject these to empirical challenges. The approach adopted in the strong program is essentially that of theory testing in science. The trouble is that this approach has limited utility in the absence of a well-developed theory to test. (Kane, 2001, p. 327)

Determining the dimensional structure of educational assessments, and then interpreting the cognitive features of those dimensions must be carried out in order to develop models of dynamic test performance. Although the special issue of *Educational Assessment* described previously is persuasive in convincing the reader of dynamic test performance in science, the validity argument can be strengthened with evidence from additional studies using different methods and with different tests (Cronbach, 1988; Kane, 2001). For example, the authors of this special issue (e.g., Roeser et al., 2002; Ayala, Shavelson, Yin, & Scultz, 2002) repeatedly cite a factor analytic study conducted by Nussbaum et al. (1997). In the study by Nussbaum et al. (1997), full-information factor analysis (implemented in the TESTFACT computer program) was used to

identify three factors underlying the science section of the NELS:88 assessment. The three factors were labeled Quantitative Science (QS), Spatial-Mechanical Reasoning (SM), and Basic Knowledge and Reasoning (BKR). Ayala, Shavelson, Yin, and Scultz (2002) replicated these results with NAEP and TIMSS using confirmatory factor analysis. However, the three factors found by Ayala et al. were highly correlated (range=.82 to .96), leading Ayala et al. to conclude “further work is needed to see whether a multiple-choice test could be constructed specifically to fit Snow’s reasoning dimensions and to test whether the dimensions do generalize, perhaps combining logical, factor analytic, and cognitive analysis in test building” (p. 118). What this preceding quote suggests is that, among other things, a substantive cognitive analysis is needed in order to better understand multidimensional science tests and why, in some cases, tests do not fit specified dimensions. A cognitive analysis of this sort is often missing, however; even in the special issue of *Educational Assessment*. An important aspect of the validity argument for dynamic test performance is the cognitive analysis of the dimensions identified. For dimensions to be informative, they must be described in detail with an eye to cognitive psychological theory. Otherwise, dimensions fail to be useful in guiding the development of dynamic models of test performance.

Educational assessments are currently not developed from cognitive models but from subject-content models (Lane, 2004; NRC, 2001). Therefore, a solid theoretical basis for specifying the number and kind of cognitive psychological factors a priori is not possible at this time and prohibits confirmatory analyses (Fabrigar et al., 1999; Preacher & MacCallum, 2003). For example, the factor descriptions presented in Nussbaum et al. (1997) are useful for categorizing items according to content but not according to cognitive processing. The factors are not described in psychological terms; at least not in sufficient detail to guide the coding of new data according to cognitive features for confirmatory analyses. Therefore, to generate hypotheses about the dynamic interplay between student and test, which then can be used by other researchers in future confirmatory analyses, exploratory analyses are a necessary first step. As Holland (1998) informs us, model building is effortful and must begin by first exploring the behavior of interest.

In an effort to begin to generate a preliminary dynamic model of science achievement, an exploratory approach was chosen for the first phase of the present study. One of the central assumptions of the dynamic view of performance is that there is an interaction between student and test, which leads to emergent patterns of behavior. These patterns of behavior first need to be identified. Currently, there are no known studies that have identified this emergent behavior for the SAIP Science Assessment. The administration of the SAIP Science Assessment, as with any large-scale science assessment, is based on the assumption that psychologically meaningful knowledge and skills are being measured in students. However, this assumption is largely untested for SAIP. Until there is some theoretical foundation for identifying the psychological factors a priori, the basis for using confirmatory analysis is weak (Fabrigar, Wegner, MacCallum, & Strahan, 1999; Preacher & MacCallum, 2003). It would also not be prudent to “just borrow” a theory about science achievement or reasoning for the purpose of conducting confirmatory analyses. It would be unclear which of the many theories that currently exist about science achievement or reasoning can or should be applied as models for the SAIP Science Assessment or any other large-scale assessment (see Holland, 1998). For example, theories about science reasoning exist from a variety of perspectives, including the developmental psychological perspective (e.g., Williams, Papierno, Makel, & Ceci, 2004), the linguistic perspective (e.g., Kelly & Bazerman, 2003), the educational psychological perspective (e.g., Echevarria, 2003), and the cognitive psychological perspective (e.g., Kuhn & Dean, 2004). There are no published criteria linking the SAIP Science Assessment to any of these theories. Furthermore, using content standards or test specifications as proxy models to guide confirmatory analyses would undermine the definition of dynamic performance because students do not respond to content standards; students react to item features that can be expected to elicit specific forms of cognitive processing. Although strong cognitive principles for confirmatory analyses are still missing in educational assessment (Cronbach & Meehl, 1955), the solution is not to use proxy models but, rather, to develop preliminary dynamic models of test

performance (Leighton, 2004; Norris, Leighton, & Phillips, 2004), which can then be subjected to further testing.

The SAIP Science Assessment

The Council of Ministers of Education in Canada (CMEC, 2000) administers the School Achievement Indicators Program (SAIP) Science Assessment, a dichotomously scored test, to students in both Grade 8 and Grade 11 (13- and 16-year-olds) every three to five years. The data gathered using the SAIP Assessment is used as a report card of Canadian students' knowledge and problem solving in science. The SAIP Science Assessment includes test items targeted to five levels of difficulty representing three broad content domains: (a) knowledge and concepts of science, (b) nature of science, and (c) relationship of science to technology and societal issues. Within the first broad content domain, *knowledge and concepts of science*, the following content knowledge associated to biology, chemistry, earth, and physics is measured: (1) matter has structure and there are interactions among its components, (2) life forms interact within environments in ways that reflect their uniqueness, diversity, genetic continuity, and changing nature, (3) basic gravitational and electromagnetic forces result in the conservation of mass, energy, momentum, and charge, and (4) earth and the physical universe exhibit form, structure, and processes of change. Within the next, broad domain, *nature of science*, the specific skill of how science involves an understanding of the nature of scientific knowledge and the processes by which that knowledge develops is measured. Within the last broad domain, *relationship of science to technology and societal issues*, the specific skill of how science involves an understanding of relationships among science, technology, and society is measured. The content domains and difficulty levels are illustrated in Table 1.

Limitations with SAIP design. The SAIP Assessment employs a two-stage testing procedure used to tailor the assessment to students' overall ability (for a review of two-stage testing procedures, the reader is referred to Bock & Zimowski, 2003). In the first stage, students are given a routing test (test A), which consists of 12 items of moderate difficulty (items targeted at level 3) used to route students to an ability-appropriate second-stage test. Dependent on students' responses to the initial 12 items, they are routed to an easier second-stage test (test B) or a more difficult second-stage test (test C). Test B consists of items targeted at low to moderate difficulty levels (1, 2, and 3), whereas test C consists of items targeted at moderate to high difficulty levels (3, 4, and 5). The first- and second-stage tests contain multiple-choice (MC) items and constructed-response (CR) items. One of the main limitations of the test data stems from the two-stage test design. To follow is a discussion of the limits the two-stage testing procedure imposes on data analysis of the SAIP Assessment.

The two-stage testing procedure introduces challenges to the types of analyses that can be conducted with the SAIP Assessment data. For instance, different proportions of 13 and 16-year-old students are routed to tests B and C. This imbalance of sample size between age groups creates difficulty when making comparisons of performance within and across age groups. Furthermore, the lack of equivalence in difficulty of items within each second-stage test imposes additional limitations in terms of the type of statistical analyses that can be conducted. Despite the richness of knowledge and skills measured by the SAIP Science Assessment, these data are probably under-utilized due to limitations the data impose on the types of analyses that can be conducted.

In accordance with the two-stage testing procedure, once students have received a score out of 12 on the routing test (test A) they are routed to either test B (where they write items 13 through 78) or test C (where they write items 79 through 144). A consequence of this two-stage procedure is that students completing either test B or C systematically fail to complete the 66 items associated with the test they did not write. In other words, there is systematic missing data in the SAIP Assessment data file for each student. Ignoring the missing data by analyzing only the complete cases is not feasible given that each observation of the data matrix unavoidably has

missing data. Particular missing data treatments such as imputation, as outlined by Rubin (1987), rely on the assumption of the randomness of the missing data pattern. Because the missing data are systematic and not random as a result of the nature of the two-stage test design, missing data treatments cannot be applied.

One method to circumvent the problem of missing data is to create two data files: for example, one file consisting of those students who wrote the routing test (test A) and test B, and another data file consisting of those students who wrote the routing test (test A) and test C. This splitting of the data file addresses the problem of missing data but does not address the problem of nonequivalent groups of 13- and 16-year-olds writing each test (AB and AC). Comparing the performance of 13-year-old students who wrote the AB test to the performance of the 13-year-olds who wrote the AC test cannot be done unequivocally because the AB and the AC tests are different tests. The same holds true for comparisons of 16-year-old students writing the AB and AC tests. Within the AB test, 13-year-olds could be compared to 16-year-olds but this comparison must be done cautiously because the sample size of 13 year-olds is greater than the sample size of 16-year-olds (the converse holds true for the AC test). The Council of Ministers of Education in Canada (CMEC) has proposed using a method to compare student performance across tests that is similar to the “vertical equating” approach presented by Zimowski, Muraki, Mislevy and Bock (1996). Vertical equating refers to the process of creating a single reporting scale, which expands over a number of student age groups in order to make comparisons across and between age groups. However, it is unclear from CMEC (2000) reports whether this has been done with the SAIP Science Assessment.

Another limitation with the SAIP Assessment data involves the inter-item correlations within tests AB and AC. Inspection of the inter-item correlations within tests suggest that the items do not correlate highly, with the highest absolute value being .20. The low inter-item correlations are a limitation resulting in part from the two-stage testing procedure. By splitting the SAIP Assessment data file into tests AB and AC (to avoid systematic missing data), the range of ability levels associated with each item is restricted, thus narrowing the variance associated with item responses. Restriction of range can attenuate correlations, and compromise the results of multivariate analyses that rely on large correlation values (e.g., low correlations can result in low factor loadings in factor analyses).

Cognitive complexity of SAIP Assessment. Despite the structural limitations with the SAIP Science Assessment described above, it represents a comprehensive measure of science achievement in students across Canada. The richness and variety of knowledge and skills measured by the SAIP Assessment provides plausible support for the multidimensional structure of science assessment. Although SAIP appears to measure multiple knowledge and skills in science comparable to NELS, NAEP, and TIMMS, these multiple dimensions still need to be identified. Currently, a single score is computed for each student’s performance on the SAIP test. If evidence is obtained to indicate that SAIP has a multidimensional latent structure, use of a single score to describe student performance masks or, even worse, fails to provide feedback to students and teachers about where students might excel and where they might require remediation. According to the dynamic view, better information can be obtained about student performance by investigating the number and nature of the latent factors that underlie a test. Upon further empirical work, these factors may be used to generate sub-scores that reflect the complexity of science achievement and the interplay between the student and test environment. There have been few studies aimed at examining the factor structure underlying the SAIP Assessment (however, see Frenette & Bertrand, 2000 for one such study). In particular, few studies (e.g., Frenette & Bertrand, 2000) have explored whether SAIP science items do indeed measure the knowledge domains outlined in the test specifications (see Table 1); that is, whether the six content domains used to develop the SAIP Assessment map onto six underlying dimensions. As educational researchers, we stand to learn a great deal about the dimensional structure of science assessments

generally and the dynamic quality of science achievement by investigating student performance on a variety of science tests, including SAIP.

Step One: Exploratory Analyses of the SAIP 1999 Science Assessment

Data from the SAIP Science Assessment administered in 1999 were used in the present study (CMEC, 2000). The AB and AC test samples from the 1999 administration were divided according to age (13- and 16-year-olds), leading to the creation of four data files—AB-13-year-olds, AB-16-year-olds, AC-13-year-olds, and AC-16-year-olds. The data files were split by age because age represents one of the most basic variable distinguishing students of science instruction—students of different ages vary in how much science instruction they have received. The four test samples were then each randomly split in half to cross validate the results obtained with each sample, thus creating eight data files as shown in Table 2, AB1-13, AB2-13, AB1-16, AB2-16, AC1-13, AC2-13, AC1-16, and AC2-16. Two exploratory techniques were used to determine the dimensional structure of the SAIP Science Assessment. These techniques are described below.¹

Exploratory DIMTEST

Dimensionality test or DIMTEST (Stout et al., 2001) has been found to be more powerful in detecting multidimensionality than other methods (see Hattie et al., 1996; Nandakumar, 1993, 1994; Nandakumar & Ackerman, 2004). DIMTEST is a nonparametric procedure used to test the null hypothesis that a set of test data is unidimensional. The null hypothesis is rejected at level α if the DIMTEST statistic, T , is larger than the 95th percentile of the standard normal distribution. The results of DIMTEST (See Table 3) indicate that for all eight data files (AB1-13, AB2-13, AB1-16, AB2-16, AC1-13, AC2-13, AC1-16, and AC2-16), the null hypothesis of unidimensionality was rejected ($p < 0.05$). (For a detailed description of DIMTEST, the reader is referred to Stout et al., 2001).

Exploratory Factor Analysis

The results from DIMTEST confirmed the multidimensional structure of the SAIP Science Assessment and revealed important information about the complexity of the test data. An exploratory factor analysis (EFA) of the tetrachoric correlations was conducted. Given the complexity of structure in the SAIP test data (see footnote 1), EFA presented a means for investigating the complex structure of certain items. EFA of the tetrachoric correlations was conducted using the program PRELIS - a preprocessor for LISREL. PRELIS estimates the Pearson correlation coefficients as tetrachoric correlation coefficients. After inputting the tetrachoric correlation matrix into SPSS an exploratory principal components analysis of the tetrachoric correlation matrix was completed using the program *Factor Analysis*. Three well-known decision rules for determining the number of factors to retain were used as well as two other decision rules more recently recommended (Preacher & MacCallum, 2003; Zwick & Velicer, 1986).

The three widely used decision rules used were: K1 (Kaiser, 1960), Cattell's Scree test (1966), and Kaiser (1958). The K1 rule (Kaiser-Guttman) suggests that the number of factors to retain is equal to the number of factors with eigenvalues equal to or greater than 1.0 (Kaiser, 1960). Cattell's Scree test (1966) suggests that if a factor is significant it will have a large eigenvalue (accounts for the majority of variance). The Scree test is a graphical approach whereby those eigenvalues that are similar in a plot will form a straight line. Eigenvalues that fall above the line are considered the factors that account for the bulk of the correlations (variance) in the matrix. The Kaiser rule for determining the number of factors to retain utilizes the Varimax-rotated image loadings matrix. In order for a factor to be retained, or considered significant, there must be at

least three factor loadings equal to or greater than 0.30. After examining all factors and factor loadings, those factors that met these criteria were retained and considered significant factors.

Beyond K1, Cattell's Scree test, and Kaiser, two additional methods for determining the number of components to retain in a factor analysis were used (Preacher & MacCallum, 2003; Zwick & Velicer, 1986).² These two additional methods include a simulation method known as parallel analysis (PA, See Horn, 1965) and a method based on the minimum average partial correlations (MAP, Zwick & Velicer, 1986). Table 4 illustrates the results of the EFA of the tetrachoric correlation matrix utilizing the five decision rules in determining the number of factors to retain. At one end of the spectrum, use of the K1 and PA rules indicated retaining over 15 factors for the AB test (13- and 16-year-olds). Use of the K1 and PA rules indicated retaining over 19 factors for the AC test (13- and 16-year-olds). At the other end of the spectrum, the Scree, Kaiser, and MAP decision rules indicated retaining between 2-3 factors for the AB-13 test and between 2-4 factors for the AB-16 test. Moreover, the Scree, Kaiser, and MAP rules indicated retaining 1-3 factors for the AC-13 test, and 2 factors for the AC-16 test. The factors retained were consistent for the cross-validation groups.

An examination of Table 4 shows a disparity in the results between PA and MAP. As mentioned by Zwick and Velicer (1986), this kind of inconsistency occurs between PA and MAP when PDCs (Poorly Defined Components) are present in the test data.³ Under these conditions, MAP is the better method to use in conjunction with the Scree test to define a range of factors. According to MAP and Scree rules, two to three factors were retained for the AB-13 test data, two factors for the AB-16 test data, two factors for the AC-13 test data, and one to three factors for the AC-16 test data.

After conducting EFA, the factors retained were rotated⁴ using orthogonal rotation procedures (i.e., quartimax, varimax) and an oblique transformation procedure (direct oblimin). We conducted both kinds of rotations/transformations to determine whether the factors should be treated as correlated or uncorrelated. One of the challenges with conducting EFA is deciding whether to interpret the orthogonal solution or the oblique solution. Orthogonal rotations treat the factors as uncorrelated, whereas oblique transformations treat the factors as correlated, normally showing minor to moderate correlations (Gorsuch, p. 188, 1983). Depending on the internal consistency of the data, different rotational methods are prescribed (Gorsuch, 1983). The internal consistency values for the SAIP data files were moderately high, with Cronbach's alpha values of .86 and .84 for the AB test for 13- and 16-year-olds, respectively; and .71 and .79 for the AC test for 13- and 16-year-olds, respectively. Gorsuch (1983, p. 185) indicates that "to apply Varimax, for example, to items of a test with high internal consistency is inappropriate because high internal consistency means there is a general factor underlying most of the items in the test." However, Michael and Bachelor (1988, p. 101-102) have demonstrated that "both an orthogonally rotated (varimax) or obliquely rotated (promax) factor solution can be expected to yield meaningful psychological dimensions" when instruments have high internal consistency.

Other investigators question the use of orthogonal rotations altogether because they doubt whether factors are ever uncorrelated in social scientific research (Fabrigar et al., 1999).⁵ For example, Fabrigar et al. (1999, p. 282) explain that researchers may prefer an orthogonal rotation because of "its simplicity and conceptual clarity (e.g., Nunnally, 1978). However, there are a number of reasons to question the wisdom of this view." Fabrigar et al. explain that for many constructs in psychology (e.g., mental abilities, personality traits, attitudes), there is substantial theoretical and empirical evidence for expecting these constructs (or dimensions of these constructs) to be correlated with one another. Therefore, they recommend conducting oblique transformations for a more accurate and realistic representation of how constructs are likely to be related to one another.

Given the above, we reasoned that it was defensible to interpret the oblique results because the factors we had identified for the SAIP data shared moderate correlations in some cases (range of .014 to .384). As mentioned previously, oblique transformations⁶ had been conducted for each of the eight data files using a range of two to three factors. Simple structure

and interpretability of the solution were two guiding principles used to determine which transformed factor solutions to retain. The criteria as set forth by Thurstone (1947) was adopted for determining if a solution approximated simple structure, which in turn would lead to a more psychologically interpretable solution (Gorsuch, 1983). After inspecting the results, we retained the (a) three-factor solution for the AB test for 13 year-olds, (b) two-factor solution for the AB test for 16 year-olds, (c) two-factor solution for the AC test for 13 year-olds, and (d) the three-factor solution for 16-year-olds. The adequacy of each of these solutions was evaluated using the substantive analyses discussed in the next section.

Tables 5 through 8 illustrate the pattern matrix for these solutions; items with loadings greater than 0.3 are shown. For each set of results (see Tables 5 through 8), the items commonly loaded on more than one factor, indicating the complexity of the data and the relative lack of simple structure. Given the nature of skills and knowledge one would expect to be measured by a science assessment, the departure from simple structure is not surprising. We turn next to understanding the nature of these skills and knowledge.

Step Two: Substantive Methods and Results

Preliminary Analysis of Transformed Factors

A common shortcoming with studies that use EFA is the sparse description of the factors found to underlie the data (Haig, 2005). Without an adequate description of the factors, researchers may not be able to replicate results with confirmatory factor analysis, understand what the dimensions are measuring in terms of skills, or confidently make inferences about students' substantive performance. An incomplete description of the factors found using EFA limits the coding of new data in future studies. As explained previously, this is one limitation we found with the NELS:88 studies (e.g., Nussbaum et al., 1997). An objective of the current study was therefore to provide a fuller description of the factors identified with EFA.

The following procedure was used to generate preliminary descriptions of the factors identified for the SAIP Science Assessment: First, items with factor loadings equal to or greater than 0.3 were retained. Then for each item with a loading equal to or greater than 0.3, the following information was recorded: (a) the first five to ten words of the test question, (b) the specific factor on which the item loaded, (c) the content standard or objective, and (d) the ability level of the item. Following a review of these item characteristics, some preliminary trends were noted. These trends are described next for the AB-13, AB-16, AC-13, and AC-16 tests collapsed across validation samples.

AB-13 test: Factor 1. Of the AB-13 items associated with factor loadings equal to or greater than 0.3 on the first factor, 29% of these items contained *key words* such as “why,” “how,” “cause/effect,” or “reason” in the test question. This clustering of items was of interest because words such as “why” or “how” are often used to evoke causal reasoning and elicit causal responses from individuals. This is supported by recent research demonstrating that specific words in problem solving tasks act as powerful cues in evoking distinct forms of reasoning (Gentner & Boroditsky, 2001; Gentner, Imai, & Boroditsky, 2002; Boroditsky, Schmidt, Phillips, 2003). Alternatively, 71% of the AB-13 items with factor loadings equal to or greater than 0.3 on the first factor contained key words such as “what,” “identify,” “sort” or “which” in the test question. Words such as “what” or “which” are normally used to evoke reasoning about class membership and cue categorical responses.

AB-13 test: Factors 2 and 3. Of the AB-13 items associated with factor loadings equal to or greater than 0.3 on the second and third factor, 10% and 7% of the test questions, respectively, contained words such as “why” or “how.” Alternatively, 90% and 92% of the items loading on the second and third factor contained words such as “what” or “which” in the test questions. Because

the items loading on the second and third factor were almost indistinguishable in their associated key words, the second and third factors were collapsed and treated as one. The mean ability level of the items loading on the first factor and second factor was 1.73 and 2.15, respectively (ability ranges from 1 to 5, with 1 being low ability and 5 being the high ability). That the ability level associated with these factors is low is not surprising given that the AB test is designed for lower-achieving students. What is interesting, however, is that the lower of the two ability levels is associated with proportionately more items eliciting causal responses (factor 1) than categorical responses. In other words, the AB test for 13-year-olds appears to measure less sophisticated causal reasoning than categorical reasoning.

AB-16 test: Factor 1. Of the AB-16 items associated with factor loadings equal to or greater than 0.3 on the first factor, 28% of these items contained key words such as “why,” “how,” “cause/effect,” or “reason” in the test question. Similar to the AB-13 test described previously, 72% of the AB-16 items with factor loadings equal to or greater than 0.3 on the first factor contained key words such as “what,” “identify,” “sort” or “which” in the test question.

AB-16 test: Factor 2. Of the AB-16 items associated with factor loadings equal to or greater than 0.3 on the second factor, 7% of the test questions contained words such as “why” or “how.” Alternatively, 93% of the items loading on the second factor contained words such as “what” or “which” in the test question. The average ability measured by the first factor was 1.828 and the average ability measured by the second factor was 1.857. Unlike the AB-13 test data, the ability levels for the two factors identified in the AB-16 test data are roughly equivalent. This suggests that the causal and categorical reasoning being measured in this test are roughly of equal complexity.

AC-13 test: Factor 1. Of the AC-13 items associated with factor loadings equal to or greater than 0.3 on the first factor, 26% of these items contained key words such as “why,” “how,” “cause/effect,” or “reason” in the test question. Alternatively, 74% of the AC-13 items with factor loadings equal to or greater than 0.3 on the first factor contained key words such as “what,” “identify,” “sort” or “which” in the test question.

AC-13 test: Factor 2. Of the AC-13 items associated with factor loadings equal to or greater than 0.3 on the second factor, 43% contained words such as “why” or “how” in the test question. Alternatively, 57% of the items loading on the second factor contained words such as “what” or “which” in the test question. The items loading on the first factor of the AC-13 test measured an average ability level of 3.76, while the items loading on the second factor measured an ability level of 4.63. The second factor, which had proportionately more items than the first factor with key words cueing causal responses, was associated with a higher ability level. In other words, the AC test for 13-year-olds appears to measure more sophisticated causal reasoning than categorical reasoning.

AC-16 test: Factor 1. Of the AC-16 items associated with factor loadings equal to or greater than 0.3 on the first factor, 15% of these items contained key words such as “why,” “how,” “cause/effect,” or “reason” in the test question. Alternatively, 85% of the AC-16 items with factor loadings equal to or greater than 0.3 on the first factor contained key words such as “what,” “identify,” “sort” or “which” in the test question.

AC-16 test: Factor 2 and 3. The items loading on the second and third factor were indistinguishable in their associated key words; therefore, the second and third factors were collapsed and treated as one. Of the AC-16 items associated with factor loadings equal to or greater than 0.3 on the second factor, 22% contained words such as “why” or “how.” Alternatively, 78% of the items loading on the second factor contained words such as “what” or “which” in the

test question. The items loading on the first factor of the AC-16 test measured an average ability level of 3.87 and the items loading on the second factor measured an average ability level of 4.23. The second factor, which had proportionately more items than the first factor with key words cueing causal responses, was associated with a higher ability level. In other words, the AC test for 16-year-olds appears to measure more sophisticated causal reasoning than categorical reasoning.

Comparison to Nussbaum et al. (1997). This preliminary description of the factors was compared to the factors identified by Nussbaum et al. (1997). As mentioned previously, following a full-information factor analyses, Nussbaum et al. identified three factors for a sample of grade 10 and grade 12 science students. The first factor identified was Quantitative Reasoning (QR), which included chemistry content and calculation or interpretation of equations, and contained items at a higher difficulty level than the other factors. The second factor identified was Spatial-Mechanical Reasoning (SMR), and included the ability to read diagrams. Finally, the third factor identified was Basic Knowledge and Reasoning (BKR), which included items requiring students to display elementary knowledge and reasoning about science concepts.

Similarly to Nussbaum et al., and other investigators (e.g., Roeser et al., 2002; Shavelson et al., 2002), we also found evidence for multiple dimensions in science assessment. However, unlike Nussbaum et al., who identified three factors, we substantively identified only two factors underlying the SAIP Science Assessment. Although both of the factors were predominately associated with test questions containing key words such as “what” or “which,” typically, one of the factors contained a higher proportion of causal-type key words than the other factor. In other words, the SAIP Science Assessment was found to contain mostly items measuring reasoning about category membership. However, some items were also found to measure reasoning about causal relationships. The most difficult version of the SAIP Science Assessment, the AC-16 test, contained the fewest items with causal key words “why” or “how.” However, these causal-type items loaded on the factor associated with a higher estimated item ability level than the factor with fewer causal-type items and lower estimated item ability level.

The factors we identified differ from Nussbaum et al.’s factors not only in number but also in description. Unlike the factors identified by Nussbaum et al., the factors we identified were not heavily anchored to content domain descriptors (e.g., chemistry, biology, or physics knowledge). The factors we identified were anchored to item features (key words) expected to yield distinct forms of cognitive processing (Gentner & Boroditsky, 2001; Gentner, Imai, & Boroditsky, 2002; Boroditsky, Schmidt, Phillips, 2003), namely causal or categorical reasoning. Although the SAIP Science Assessment includes six content domains, these content domains were evenly represented across the two factors. Therefore, generating content descriptions of the factors was difficult and would have proven to be uninformative. Rather, the factors we identified were anchored more heavily to the hypothesized cognitive processes likely to be invoked by the items in students irrespective of content—such as causal or categorical reasoning. There is one similarity in the factors found in the current study and in the Nussbaum et al. study. The BKR factor found in the Nussbaum study reflected general reasoning about science, which can be assumed to combine different forms of reasoning such as reasoning about causes and effects, and category membership. The QR and SMR factors found by Nussbaum et al. were not easily matched to the items in the SAIP Assessment. This inability to match might have occurred because identification of the item features associated with the QR and SMR factors was not explicit in the Nussbaum et al. study. Alternatively, unlike the SAIP Science assessment, the NELS:88 for 10th and 12th grade students may incorporate a wider array of skills, facilitating the identification of factors such as QR and SMR in addition to BKR. It is possible that the SAIP Assessment only measures knowledge and skills reflected in the BKR factor, which in our study, revealed itself as two factors—one measuring proportionately more causal reasoning than the other, which measures more categorical reasoning.

To describe student performance as dynamic, however, entails more than identifying the similarities associated with a set of test items. Attempts must be made to describe the test item

features that are believed to elicit distinct forms of thinking. Then, an explanation should be given as to why these forms of thinking are distinct. Without this explanatory effort, we are left with assigning a label to a factor and assuming, incorrectly, that the label somehow explains the behavior (see nominal fallacy, Levy, 1997); and we acknowledge that to name the behavior does not explain it. Our preliminary factor descriptions are subject to the same critique. Our two identified factors—reasoning about causes and effects and reasoning about categories—are only labels that still require a fuller description. Describing a model of why these factors might elicit distinctive behaviors from students begins the process of generating testable hypotheses about student test performance (see Holland, 1998). In the next section, we describe a theory of scientific reasoning that involves causal and associative/categorical processes. We review this theory because it (a) represents a dynamic theory of scientific reasoning, (b) clarifies the factors found for the SAIP Science Assessment, including the details of item characteristics that can be coded and organized for confirmatory analyses, and (c) provides a theoretical basis for the validity argument for why causal reasoning and categorical reasoning are useful descriptors of science performance.

A Dynamic Theory of Scientific Reasoning

Our preliminary analyses of the AB and AC tests suggested that the two factors underlying the SAIP Science Assessment tapped student reasoning about causes and effects and student reasoning about category membership. However, naming these factors did not explain them or give us any information about why students would respond differentially to questions requesting causal answers and questions requesting categorical answers. We therefore turned to the cognitive developmental literature for a scientific reasoning theory that might illuminate the psychological reasons for students' responses.

An examination of theories of scientific reasoning led to a recently published review article (2004) by Deanna Kuhn and David Dean Jr.. Deanna Kuhn is a leading authority on inductive reasoning, and scientific reasoning in particular. Kuhn has devoted her career to investigating the psychological processes of individuals as they coordinate theory and evidence, and the contextual variables associated with causal and non-causal inferences (Kuhn, 2001). In their article titled "Connecting Scientific Reasoning and Causal Inference," Kuhn and Dean review the bodies of literature associated with research on multivariable causal inference and scientific reasoning in students and adults. They succinctly summarize the objective of scientific reasoning: "in science, a major goal is to reduce complexity to manageable levels and thereby enhance explanatory power, by at least temporarily eliminating factors from consideration" (p. 266).

In the ongoing process of managing and reducing the complexity of information from the external environment, Kuhn and Dean explain that individuals typically make use of two forms of inference—causal and non-causal. Causal inference is directed and satisfies the purpose of explaining phenomena. Non-causal inference can be viewed as associative and satisfies the purpose of categorizing phenomena. Students' beliefs about causal mechanisms influence the task variables believed to be worthy of further exploration and analysis. A task variable that is believed to have a causal effect is expected to have a temporal relationship to an effect or outcome of interest; therefore causal claims are often subjected to a higher standard of evidence and explanation than non-causal (i.e., categorical) claims. The need to mentally represent and find empirical support for the temporal relationship in causal inference raises the evidential standard of causal inference (Klaczynski, 2000; Kuhn, 1995). In contrast, a task variable believed to be only associated with an outcome can be categorized with the outcome simply by noting the covariation between the variable and outcome. In this case, the variable does not need to have a temporal relationship with the outcome to say that it can be categorized with the outcome. Such a variable might be categorized with a phenomena of interest but is not necessarily considered explanatory of the phenomena. Causal inferences in particular influence how students interpret new data and

generate predictions. Kuhn and Dean (2004, p. 273) emphasize the development of novice scientific reasoning to more sophisticated reasoning:

Another notable finding, common to all age groups, is the asymmetry between causal and noncausal inference. Modification of initial beliefs from causal to noncausal appears to be more challenging than is change in the other direction, that is, from a noncausal to a causal belief. In the latter case, individuals are likely to resist interpreting covariation data suggesting a causal effect until they have constructed a theory to explain the effect, but once they have devised one, they show themselves able and willing to modify an initial noncausal judgment to an inference of causality justified by covariation data.

Kuhn and Dean further explain that individuals are often inconsistent in their efforts to coordinate claims about theory and evidence. Individuals are inconsistent because the inferences they generate—causal or non-causal—depend on whether they are accurately representing all the variables in their mental model of the problem. Moreover, given distinct contexts, individuals may view a variable as sometimes causal and sometimes not. Depending on the context, the evidential standards for judging a variable as causal can change. For example, consider a statement “If Mary eats candy, then she will get cavities.” Few individuals would query the causal mechanism linking eating candy to cavities if it is also known that Mary *does not* brush her teeth. However, if no information is provided about Mary’s dental hygiene, then individuals might be more ready to assume that candy eating will not lead to cavities because a number of possible situations exist for circumventing the cavities such as daily brushing or frequent visits to the dentist (see Cummins, 1995, for additional examples). Inconsistent inferences about cause and effect can, in turn, create variation in the predictions made about an outcome of interest. Causal inferences are almost always justified by reference to temporal relationships, whereas non-causal inferences invariably lack this standard of evidence. In sum, Kuhn and Dean (2004, italics added, p. 285) emphasize that:

Theory-evidence coordination is *a complex, dynamic process*, with the role of theory not confined to an initial phase in which variables are excluded from consideration on theoretical grounds. The data presented here contain frequent instances of an individual’s shift from an earlier declaration of a variable as noncausal to a subsequent claim that it is causal...Rather than seeking to identify a universal set of rules that characterize causal inference within and across individuals, an objective debated recently by others (Nisbett, Peng, Choi, & Norenzayan, 2001), we propose as a more accurate working model one in which an individual brings a set of varying inference strategies, or rules (of varying validity), to the task of interpreting the implications of evidence...”

Using causal and non-causal (i.e., categorical) inference as a backdrop, we reviewed the items of the SAIP Science Assessment and three raters coded all items that loaded in the EFA according to whether the test questions contained primarily causal or categorical-type key words. In our coding of items, we used key introductory words such as “why,” “how,” “cause/effect,” “what,” “which,” or “identify” to code items as either *primarily causal* or *primarily categorical*. Three raters coded the items and 90% inter-rater agreement was obtained.

Coding the items according to whether they contained causal key words (e.g., why or how) or categorical key words (e.g., what or which) was on the surface straightforward because the focus was on key words alone. However, there was a deeper distinction to be made between underlying causal and categorical themes. Separating causal and categorical themes was more difficult to discern than originally anticipated. Consider that an item might contain a causal key word but essentially be requesting a categorical response. For example, the multiple choice (MC) question,

“How were these mountains formed?” (item 30) was coded as a primarily causal item because this question contains the causal key word “how.” However, the format of the MC item presents a delimited context to students. This delimited context may evoke students to create a representation of the item that is more aligned with generating a categorical inference than a causal inference. Only four alternatives are provided as possible answers, and each answer represents a single phrase *without any explication of the temporal relationship underlying the formation of mountains*. You might reasonably expect students responding to this MC item to interpret it as essentially requesting a categorical response rather than a causal response.

Next, consider the following constructed-response (CR) question, “What could they do to identify the animal tracks in the snow?” (item 33). This question contains the categorical key word “what” and therefore was coded as primarily categorical. However, after discussing this item, it was determined that this question is essentially requesting students to provide a plan for identifying animal tracks in the snow; the demand for how the plan is to be used to identify tracks is *essentially a request for a temporal mechanism* to achieve a specific result. Given the possible influence of format on students’ mental models of the item (i.e., whether the question is interpreted as asking for a causal response or a categorical response), we also coded SAIP items according to their format. We reasoned that format might function as a proxy for invoking either causal or categorical reasoning. Following the coding of items, confirmatory analyses were conducted and are described in the next section.

Confirmatory Analyses of Causal-Categorical Model (CCM), Item Format Model (IFM) and Test Specifications Model (TSM)

We conducted a linear factor analysis with LISREL⁶ to estimate the parameters for a 2-dimensional model using item coding associated with the Causal-Categorical Model (CCM) and the Item Format Model (IFM). We also used a linear factor analysis to estimate the parameters for a 6-dimensional model using item coding associated with the Test Specifications Model (TSM).^{7,8}

The results obtained from the analyses for all three models are presented in Tables 9, 10, and 11. The right hand panel of Tables 9, 10, and 11 illustrates three indices (i.e., RMSEA, RMR, and AGFI) for judging the model-data fit of the CCM, IFM, and TSM. The RMSEA provides a measure of parsimony by assessing the number of free parameters required to achieve a given level of fit. Browne and Cudeck (1993) recommend that an RMSEA of .08 to .05 indicate a close fit of the model in relation to the degrees of freedom. The RMR is the root mean of squared discrepancies between the observed covariances fitted and the hypothesized covariances. RMR provides a goodness-of-fit measure, where a small RMR (.05) indicates good fit. Finally the AGFI index is a goodness of fit index adjusted for degrees of freedom. Values of AGFI greater than .9 indicate excellent fit, while values greater than or equal to .8 indicate good fit. A fourth index is the chi-square statistic but it is not reported here because it is not a useful index with very large sample sizes. According to Gierl and Rogers (1996), the chi-square statistic is not a helpful index by which to assess model fit because the statistic is “dependent on sample size and often results in a statistically significant difference when large samples are used, even when fit appears good using other indexes” (p. 319).

Following the guidelines suggested by Gierl and Rogers (1996) and Browne and Cudeck (1993), the fit indices for all three models, CCM, IFM, and TSM, were evaluated. For the AB-13 test data, the value of the RMSEA was identical for all three models (.057). The value of RMR was slightly better for the IFM and CCM (.055), than for the TSM (.056). The value of AGFI was .79 for all three models. These values suggest that for the AB-13 test data, all three models provide an equally adequate fit. For the AB-16 test data, the value of the RMSEA was slightly better for the IFM (.073) than for the CCM and TSM (.074). The value of RMR was identical for all three models (.066). The value of AGFI was slightly better for the IFM and TSM (.76) than for the CCM (.75). These values suggest that for the AB-16 test data, all three models provide an equally adequate fit. Although none of the models fit the AB test data poorly, the models only fit

adequately. One reason for this outcome might be the nature of the student sample writing the AB test. There were many incomplete cases that were excluded from the analysis. This is not uncommon according to the Council of Ministers of Education (CMEC), given that low-achieving students often miss series of questions or refuse to continue with the exam.

For the AC-13 test data, the value of the RMSEA was lowest for the IFM (.048), with slightly higher values for the CCM (.051) and TSM (.050). The value of RMR was lowest again for the IFM (.041), but identical for the CCM and TSM (.044). The values of AGFI were highest for the IFM (.92) and slightly lower for the CCM (.90) and TSM (.91). These values suggest that for the AC-13 test data, the IFM provides a consistently better fit than the CCM and TSM. For the AC-16 test data, the value of the RMSEA was lowest for the IFM (.042), with higher values for the CCM (.047) and TSM (.044). The value of RMR was lowest for the IFM (.038), with higher values for the CCM (.042) and TSM (.040). The value of AGFI was highest for the IFM (.93), with lower values for the CCM (.91) and TSM (.92). These values suggest that for the AC-16 test data, the IFM provides a consistently better fit than the CCM and TSM. The best fitting model for the AC test data was the IFM.

General Discussion

The goal of the present study was to identify the latent dimensional structure of a new large-scale science assessment called the School Achievement Indicators Program (SAIP) Science Assessment. The first objective of the current study was to use DIMTEST (Stout, Douglas, Junker, & Roussos, 1993), in addition to exploratory factor analysis (Zwick & Velicer, 1986), to determine the dimensional structure of the SAIP Science 1999 Assessment. Moreover, if we found more than one factor to underlie the data, the second objective was to identify and describe the substantive cognitive features of the factors to provide more informative, relevant information about the potential processes underlying student performance on science assessments.

In response to the first objective, our statistical results indicate evidence for the multidimensional latent structure of the SAIP Science Assessment. DIMTEST and EFA results all indicated that the SAIP Science Assessment measures more than one dimension. Our results provide further empirical support for the research conducted by other investigators, who have also found evidence for the multidimensional nature of science assessment (e.g., Roeser et al., 2002; Shavelson et al., 2002). In particular, DIMTEST confirmed that there is more than one dimension measured by SAIP, while EFA confirmed that approximately two dimensions are being measured.

We found evidence for fewer factors than Nussbaum et al.'s (1997) study of the NELS: 88 test data. In the Nussbaum et al. (1997) study, the investigators found three factors to underlie the data: Background knowledge and Reasoning (BKR), Spatial-Mechanical Reasoning (SMR), and Quantitative Reasoning (QR). Our results indicated evidence of the first factor (BKR) but not the SMR and QR factors. A possible reason for this is the variation in the development and design of the different tests. The NELS:88 test items may be designed to measure a larger variety of content skills. In contrast, the SAIP Science Assessment generally contained questions targeted to the BKR factor, which in the present study was identified as consisting of two factors—a causal reasoning factor and a non-causal (categorical) reasoning factor.

In response to our second objective, we attempted to identify the psychological characteristics of the factors we found to underlie the data. Providing a description of the psychological characteristics was necessary to move closer to a dynamic description of student performance. According to the dynamic perspective, the focus is on emerging intelligent behavior (Pfeifer & Scheier, 1999, p. 632):

We expect from a theory of intelligence an answer to the following question: Given an agent, an animal, or a human, that exhibits certain behaviors, what are the underlying mechanisms? ... This question is of central importance to the study of intelligence: It pertains to diversity-compliance considerations that we identified... as being a core

characteristic of intelligence. Diversity-compliance refers to a trade-off that any intelligent agent must resolve, a trade-off between a conservative aspect that exploits givens, and one that is responsible for generating [within itself] the diversity required to remain adaptive.

This was not an easy task given the relative dearth of information about the cognitive processes and strategies students apply to interpret and solve science test questions. To achieve this objective we categorized the items based on key words such as “What” “Why” or “How.” Key words such as these have psychological effects, cuing distinct forms of reasoning and responses (Gentner & Boroditsky, 2001). Framed within a science context, these key words might be interpreted in specific ways by students, leading them to react differently to classes of items. We hypothesized that items anchored to “why” or “how” words might cue causal responses and items anchored to “what” or “which” words might cue categorical responses. Kuhn and Dean’s (2004) recent literature review of multivariable and scientific reasoning provided support of our hypothesis. In this review, these investigators claim that empirical research demonstrates that individuals generate causal inferences and non-causal (i.e., categorical) inferences using different evidential standards. For individuals to make causal inferences, typically, some mechanism must be identified in order to temporally link variables. This is not required for categorical inference. In order to make a categorical inference, no mechanism or theory is needed to link variables temporally other than the perceived covariation of variables. Given the substantive support this paper provided to the preliminary factors we identified, items were coded according to the Causal-Categorical Model (CCM). Items were also coded according to item format (IFM)—multiple choice or constructed-response. Constructed-response questions through their open format may cue causal type responses because students are expected to extend their descriptions by providing temporal information about sequences of action. The item format model (IFM), along with CCM and test specifications model (TSM), was tested using confirmatory factor analysis. Results from these analyses indicated that the IFM model was the better fit to the AC test data. The results were generally less clear for the AB test data because a large number of cases were deemed invalid.

The superiority of the IFM suggests that item format is an important variable in eliciting distinct forms of reasoning. Although the constructed-response format has been previously connected with higher levels of thinking in students, the empirical research supporting this connection has been mixed (e.g., Resnick & Resnick, 1990; Rodriguez, 2002). Different formats, for example, constructed response (CR) are consistently implemented in large-scale assessments because of a belief that they offer a way of measuring higher-level reasoning skills, which may not be possible with multiple-choice items (MC) (Resnick & Resnick, 1990; Rodriguez, 2002). However, when the goal is to measure complex cognitive functions, the research suggests that MC and CR items can measure similar constructs and levels of reasoning (Katz, Bennett, & Berger, 2000; Rodriguez, 2002). That MC items elicit factual knowledge and recognition is a criticism not of item format but directly related to the way items are written (Haladyna, 1999, 2004). Furthermore, no research has been undertaken to critically investigate the psychological processes that are responsible for the advantage normally assigned to the constructed-response format. Our research suggests that one reason why the constructed-response format may elicit higher-level thinking in students is because students are more likely to engage in causal reasoning. According to Kuhn and Dean’s (2004) review of scientific reasoning, individuals explain the external environment using two forms of reasoning. Causal inferences are made when individuals can generate temporal connections between variables, often in the form of a temporal mechanism. Categorical inferences are made when individuals perceive variables as co-varying together but no attempt is made to link the variables via a temporal mechanism. The standard of evidence is therefore higher for causal inferences than for categorical inferences. If students in science perceive constructed-response items as requests for information detailing some mechanism which links variables, then students might generate and evaluate their responses using a higher standard. If, moreover, students perceive multiple-choice questions as requests for categorical information

alone, then students might use a lower standard for not only generating their responses but also evaluating them. Our results provide another, theoretically- based perspective for investigating the cognitive processes of students as they respond to MC or CR science questions.

The present study indicates that science assessments are multidimensional. Two forms of reasoning prove useful in describing the dimensions underlying the SAIP Science Assessment—causal and categorical. As Kuhn and Dean (2004) explain, reasoning about causes demands a higher standard of theory-evidence coordination than reasoning about categories. It would be tempting at this stage to argue that science assessments are perhaps essentially unidimensional because they represent primarily measures of causal reasoning. Although tempting to generate such a bold conclusion, it would not be substantiated by the results we have obtained. Although causal inference may require a higher level of evidence from students, this does not suggest that categorical reasoning is unimportant. Causal and categorical reasoning are both meaningful reasoning dimensions, which demand from students distinct standards of evidence. The former involves identifying a temporal mechanism by which two variables connect, while the latter involves identifying that two variables simply covary. To be sure, these two forms of reasoning are related in their focus to reduce environmental complexity, and their goal to generate better inferences with which to understand the world. However, each nonetheless demands a distinct standard of evidence. Hence, when we describe science assessments as multidimensional, we may in fact be describing two forms of reasoning that individuals use to manage the complexity of environmental variables. We are currently investigating the cognitive processes students use to answer SAIP science items.

The present study also finds support for the dynamic nature of science achievement—and possibly all achievement. Students interpret the cognitive demands of questions and respond using distinct standards of evidence. Finally, the main advantage of retrofitting cognitive models to test data may be the identification of *hypotheses* about what students react and respond to in the tests they write. The results from this study can be viewed as yet further evidence that tests need to be developed from explicit and empirically-based cognitive models if we are to have more than hypotheses and be fully confident in what tests measure in students.

Footnotes

¹DETECT, a nonparametric statistical procedure that uses conditional covariance to partition test items into cohesive, mutually exclusive clusters, was also used (Zhang & Stout, 1996). When simple structure is not found in the data, DETECT does not reliably partition the item clusters. For this reason, the DETECT results were not used in the present study.

²When conducting EFA, one of the critical decisions that need to be made is the number of factors to retain in representing the factor structure of the data. In answer to this decision, Zwick and Velicer (1986) conducted a simulation study where they used principal component analysis and five different decision rules, Bartlett's chi-square test, K1, MAP, scree test, and PA with data of known structure and complexity. In their study, they focused on the ability of these five decision rules to estimate the number of components in the population correlation matrices given sample matrices. Zwick and Velicer concluded that scree method was more accurate than the K1 method and Bartlett's test, but tended to overestimate the number of factors to retain at low levels of saturation. Zwick and Velicer suggested using the scree test with either MAP or PA to decide on a range of factors. Other researchers also recommend using multiple decision rules, and in particular PA or MAP, to decide on the range of factors to retain (Fabrigar, Wegner, MacCallum, & Strahan, 1999; Haig, 2004; Preacher & MacCallum, 2003). According to Zwick and Velicer (1986), the two most accurate methods were MAP and PA but these decision rules led to divergent results when poorly defined components (PDCs) were present in the data.

³Poorly defined components are those that involve a small number of variables or have low saturation. PA retains poorly defined components and therefore overestimates the number of components to retain in data containing PDCs, but MAP does not. Consequently, the use of MAP with data containing PDCs is recommended.

⁴In cases where a range of factors was retained, rotations were done on each set of factors.

⁵However, Gorsuch (1983, p. 188) indicates that "To rotate orthogonally it must be assumed that the factors are uncorrelated. Many investigators would prefer either to test this assumption or to allow some minor correlation among the factors. ... The degree of correlation allowed among factors in an oblique rotation is always minor to moderate and can be influenced by user defined parameters in several procedures. No one allows the factors to become highly correlated; if two factors ever did become highly correlated, most investigators would redo the analysis with one less factor."

⁶Oblique transformation was conducted with a $d=0$. The d or delta is a user-controlled parameter that controls the extent of obliqueness among the factors. Negative values of d decrease the correlation among factors, whereas positive values of d increase the correlation. A d of zero is the default. Given this that parameter "is generally left exclusively to the user so that [it] is a guess" (Gorsuch, p.189, 1983), a default of zero was considered to be the least contentious choice given the exploratory nature of the analysis.

⁷Although linear factor analysis is a commonly used procedure, Embretson and Reise (2000) have recommended researchers move away from using commonplace heuristic procedures for determining dimensional structures such as proportion of variance accounted for by the first factor or ratio of first to second eigenvalue. Embretson and Reise (2000) recommend greater use of other techniques such as those found in TESTFACT, POLYFACT, NORHARM, and LISCOMP. It is understandable that Embretson and Reise (2000) would make such a recommendation given that many of the behaviours psychometricians are interested in modeling may involve non-linear relationships. In fact, Pfeifer and Scheier (1999, p. 632) suggest that nonlinear techniques are often necessary to model dynamic systems. Consequently, in addition to conducting a linear factor analysis, we also conducted a non-linear factor analysis with NOHARM (normal ogive by harmonic analysis robust method; Fraser, 1988; McDonald, 1967, 1999) to estimate the parameters for the models using items associated with the CCM, IFM, and TSM codings. The results from the non-linear factor analysis were similar to the results from the linear factor analysis. However, we focus on the LISREL results in the current paper instead of the

NOHARM results because the former has clearer guidelines for interpreting model-data fit indices.

⁸Although the AB and AC test data were found to exhibit complex structure according to EFA results, the IFM, CCM and TSM models were created to exhibit simple structure. Simple structure provides the most parsimonious representation of the data. Furthermore, modeling the data using models exhibiting complex structure led to poorer results.

References

- Ackerman, T.A., Gierl, M.J., & Walker, C.M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, Fall*, 37-53.
- Ayala, C.C., Shavelson, R.J., Yin, Y., & Schultz, S.E. (2002). Reasoning dimensions underlying science achievement: The case of performance assessment. *Educational Assessment*, 8, 101-121.
- Bartholomew, D. J. (1987). *Latent variable models and factor analysis*. London, England: Griffin.
- Bernstein, I. H., & Teng, G. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin*, 105, 467-477.
- Bock, R. D., & Zimowski, M. (2003). NAEP validity studies: Feasibility studies of two-stage testing in large-scale educational assessment: Implications for NAEP, NCES 2003-14, by R. Darrel Bock and Michelle Zimowski. U.S. Department of Education, National Center for Education Statistics, Washington, DC.
- Boroditsky, L., Schmidt, L.A., & Phillips, W. (2003) Sex, syntax and semantics. In A.C's *Language in mind: Advances in the study of language and thought*. (pp. 61-79). Cambridge, MA, US: MIT Press.
- Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245-276.
- Charness, N., & Schultetus, R. S. (1999). Knowledge and expertise. In F. T. Durso, R. S., Nickerson, R. W. Schvaneveldt, S. T. Dumais, D. S. Lindsay, and M. T. H. Chi (Eds.), *Handbook of applied cognition* (pp. 57-81). Wiley.
- Cota, A. A., Longman, R. S., Holden, R. R., Fekken, G. C., & Xinaris, S. (1993). Interpolating 95th percentile eigenvalues from random data: An empirical example. *Educational and Psychological Measurement*, 53, 583-596.
- Council Ministers of Education, Canada (2000). *Public report on science assessment: SAIP School Achievement Indicators Program 1999*. Retrieved August 12, 2002, from <http://www.cmec.ca/saip/science2/science2.en.stm>.
- Cronbach, L.J., & Meehl, P. E. (1955). *Psychological Bulletin*, 52, 281-302.
- Cronbach, L.J. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Erlbaum.
- Cummins, D.D. (1995). Naïve theories and causal deduction. *Memory & Cognition*, 23, 646-658.
- Dixon, R.A. (1992). Contextual approaches to adult intellectual development. In R.J. Sternberg and C.A. Berg (Eds.), *Intellectual development* (pp. 350-380). Cambridge: Cambridge University Press.
- Douglas, J., Kim, H., Roussos, L., Stout, W., & Zhang, J. (1999). *LSAT dimensionality analysis for the December 1991, June 1992, and October 1992*. LSAC Research Report Series. Law School Admission Council, Inc.
- Echevarria, M. (2003). Anomalies as a catalyst for middle school students' knowledge construction and scientific reasoning during science inquiry. *Journal of Educational Psychology*, 95, 357-374.
- Embretson, S.E., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
- Fabrigar, L.R., Wegener, D.T., MacCallum, R.C., & Strhan, E.J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4, 272-299.
- Fischer, K.W., & Pipp, S.L. (1984). Processes of cognitive development: Optimal level and skill acquisition. In R.J. Sternberg (Ed.), *Mechanisms of cognitive development* (pp. 45-80). New York: Freeman.
- Fraser, C. (1988). *NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory*. Armidale, Australia: The University of New England.

- Frenette, E. & Bertrand, R. (2000, April). *Assessing dimensionality with TESTFACT and DIMTEST using large-scale assessment data sets*. Paper presented at the annual meeting of the American Educational Research Association (AERA). New Orleans, LA.
- Froelich, A.G. (2000). *Assessing the unidimensionality of test items and some asymptotics of parametric item response theory*. Unpublished Doctoral Dissertation. University of Illinois at Urbana-Champaign, Department of Statistics.
- Gardner, H. (1993). *Multiple intelligences. The theory in practice*. New York: Basic Books.
- Gentner, D., & Boroditsky, L. (2001). Individuation, relational relativity and early word learning. In M. Bowerman and S. Levinson (Eds.), *Language acquisition and conceptual development*. Cambridge, UK: Cambridge University Press.
- Gentner, D., Imai, M., & Boroditsky, L. (2002). As time goes by: Evidence for two systems in processing space time metaphors. *Language & Cognitive Processes*, 17, 537-565.
- Gierl, M.J., & Rogers, W.T. (1996). A confirmatory factor analysis of the test anxiety inventory using Canadian high school students. *Educational and Psychological Measurement*, 56, 315-324.
- Gierl, M. J., Leighton, J.P., & Tan, X. (May, 2005). *Evaluating the Cluster Consistency of DETECT When Data Display Complex Structure*. Paper presented at the annual meeting of the Canadian Society for Studies in Education (CSSE), London, Ontario, Canada.
- Giroto, V. (2004). Task understanding. In J. P. Leighton and R. J. Sternberg (Eds.), *Nature of reasoning* (pp. 103-128). NY: Cambridge University Press.
- Glorfeld, L. W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement*, 55, 377-393.
- Gorsuch, R. L. (1983). *Factor analysis (2nd Ed.)*. Hillsdale, NJ: Erlbaum.
- Haig, B.D. (2005). *Exploratory factor analysis, theory generation, and the scientific method*. Manuscript under review
- Haladyna, T.M. (1999). *Developing and validating multiple-choice test items (2nd ed.)*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.
- Haladyna, T.M. (2004). *Developing and validating multiple-choice test items (3rd ed.)*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.
- Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: SAGE Publications.
- Hamilton, L.S., Nussbaum, E. M., & Snow, R. E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education*, 10, 181-200.
- Hattie, J., Krakowski, K., Rogers, H.J., Swaminathan, H. (1996). As assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement*, 20, 1-14.
- Haydel, A.M., & Roeser, R.W. (2002). On motivation, ability, and the perceived situation in science test performance: A person-centered approach with high school students. *Educational Assessment*, 8, 163-189.
- Holland, J.H. (1998). *Emergence: From Chaos to Order*. Reading, MA: Addison-Wesley.
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179-185.
- Jöreskog, K. G., & Sörbom, D. (1996). *PRELIS 2: User's reference guide*. Chicago, IL: Scientific Software.
- Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20, 141-151.
- Kane, M.T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38, 319-342.
- Katz, I.R., Bennett, R.E., & Berger, A.E. (2000). Effects of response format on difficulty of SAT-mathematics items: It's not the strategy. *Journal of Educational Measurement*, 37 (1), 39-57.
- Kelly, G. J., & Bazerman, C. (2003). How students argue scientific claims: A rhetorical-semantic analysis. *Applied Linguistics*, 24, 28-55.
- Kim, H.R. (1994). *New techniques for the dimensionality assessment of standardized test data*. Unpublished Doctoral Dissertation, University of Illinois at Urbana-Champaign, Champaign.

- Klaczynski, P. (2000). Motivated scientific reasoning biases, epistemological beliefs, and theory polarization: A two-process approach to adolescent cognition. *Child Development, 71*, 1347-1366.
- Kuhn, D. (1995). Microgenetic study of change: What has it told us? *Psychological Science, 6*, 133-139.
- Kuhn, D. (2001). Why development does (and does not occur) occur: Evidence from the domain of inductive reasoning. In J. L. McClelland & R. Siegler (Eds.), *Mechanisms of cognitive development: Behavioral and neural perspectives* (pp. 221-249). Hillsdale, NJ: Erlbaum.
- Kuhn, D., & Dean Jr., D. (2004). Connecting scientific reasoning and causal inference. *Journal of Cognition and Development, 5*, 261-288.
- Kuppermintz, H. (2002). Affective and conative factors as aptitude resources in high school science achievement. *Educational Assessment, 8*, 123-127.
- Lane, S. (2004). 2004 NCME Presidential Address. Validity of high-stakes assessment: Are students engaged in complex thinking? *Educational Measurement: Issues and Practice, 23*, 6-14.
- Lau, S., & Roeser, R. W. (2002). Cognitive abilities and motivational processes in high school students' situational engagement and achievement in science. *Educational Assessment, 8*, 139-162.
- Leighton, J. P. (2004). Avoiding Misconceptions, Misuse, and Missed Opportunities: The Collection of Verbal Reports in Educational Achievement Testing. *Educational Measurement: Issues and Practice, 23*, 1-10.
- Leighton, J. P., Gierl, M. J., & Hunka, S. (2004). The attribute hierarchy model: An approach for integrating cognitive theory with assessment practice. *Journal of Educational Measurement, 41*, 205-236.
- Leighton, J. P., & Sternberg, R. J. (2003). Reasoning and problem solving. In A. F. Healy & R. W. Proctor (Volume Eds.), *Experimental Psychology* (pp. 623-648). Volume 4 in I. B. Weiner (Editor-in-Chief) *Handbook of psychology*. New York: Wiley.
- Levy, D.A. (1997). *Tools of critical thinking: Metathoughts for psychology*. Boston: Allyn and Bacon.
- McDonald, R.P. (1967). Nonlinear factor analysis. *Psychometric Monographs, No. 15*.
- McDonald, R.P. (1999). *Test theory: A unified approach*. Mahwah, NJ: Erlbaum.
- Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11*, 3-31.
- Nandakumar, R. (1993). Assessing essential unidimensionality of real data. *Applied Psychological Measurement, 1*, 29-38.
- Nandakumar, R. (1994). Assessing latent trait unidimensionality of a set of items: Comparison of different approaches. *Journal of Educational Measurement, 31*, 1-18.
- Nandakumar, R., & Ackerman, A.A. (2004). Test modeling. In K. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social science* (pp. 93-105). Thousand Oaks, CA: Sage.
- Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics, 18*, 41-68.
- National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
- Norris, S.P., Leighton, J.P., & Phillips, L.M. (2004). What is at stake in knowing the content and capabilities of children's minds? A case for basing high stakes tests on cognitive models. *Theory and Research in Education, 2*, 283-308.
- Nussbaum, E.M., Hamilton, L.S., & Snow, R. E. (1997). Enhancing the validity and usefulness of large-scale educational assessments: IV. NELS:88 Science Achievement to 12th Grade. *American Educational Research Journal, 34*, 151-173.
- Pfeifer, R., & Scheier, C. (1999). *Understanding intelligence*. Cambridge, MA: The MIT Press.
- Preacher, K.J., & MacCallum, R.C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics, 2*, 13-43.
- Puhan, G. (2003). Evaluating the effectiveness of two-stage testing for English and French examinees on the SAIP Science 1996 and 1999 tests.

- Resnick, L.B., & Resnick, D.P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B.R. Gifford & M.C. O'Connor (Eds.), *Changing Assessments: Alternative Views of Aptitude, Achievement and Instruction*. Norwell, Massachusetts: Kluwer Academic Publishers.
- Rodriguez, M.C. (2002). Choosing an item format. In G. Tindal & T.M. Haldyna (Eds.). *Large-Scale assessment programs for all students: Validity, technical adequacy, and implementation*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.
- Roeser, R.W., Shavelson, R.J., Kupermintz, H., Lau, S., Ayala, C., Haydel, A., Schultz, S., Gallagher, L., & Quihuis, G. (2002). The concept of aptitude and multidimensional validity revisited. *Educational Assessment, 8*, 191-205.
- Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons.
- Shavelson R.J., Roeser, R. W., Kupermintz, H., Lau, S., Ayala, C., Haydel, A., Schultz, S., Gallagher, L., & Quihuis, G. (2002). Richard E. Snow's remaking of the concept of aptitude and multidimensional test validity: Introduction to the special issue. *Educational Assessment, 8*, 77-99.
- Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263-331). New York: American Council on Education, Macmillan.
- Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52*, 589-617.
- Stout, W. F., Douglas, J., Junker, B., & Roussos, L.A. (1993). *DIMTEST manual*. Unpublished manuscript available from W. F. Stout, University of Illinois at Urbana-Champaign, Champaign.
- Stout, W.F., Habing, B., Douglas, J., Kim, H.R., Roussos, L.A., & Zhang, J. (1996). Conditional covariance based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20*, 331-354.
- Stout, W., Froelich, A. G., & Gao, F. (2001). Using resampling to produce an improved DIMTEST procedure. In A. Boomsma, M.A.J. van Duijn, & T.A.B. Snijders (Eds.), *Essays on item response theory* (pp. 357-376). NY: Springer-Verlag.
- Thagard, P. (1996). *Mind: Introduction to Cognitive Science*. MIT Press.
- Thurstone, L.L. (1947). *Multiple-factor analysis: a development and expansion of The vectors of mind*. Chicago: University of Chicago Press.
- Van Geert, P. (2003). Measuring intelligence in a dynamic systems and contextualist framework. In R.J. Sternberg, J. Lautrey, and T.I. Lubart (Eds.), *Models of intelligence: International perspectives* (pp. 195-211). Washington DC: American Psychological Association.
- Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41*, 321-327.
- Williams, W., Papierno, P. B., Makel, M. C., Ceci, S. J. (2004). Thinking Like A Scientist About Real-World Problems: The Cornell Institute for Research on Children Science Education Program. *Journal of Applied Developmental Psychology, 25*, 107-126.
- Zhang, J., & Stout, W.F. (1996, April). *A new theoretical DETECT index of dimensionality and its estimation*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.
- Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG Multiple-group IRT analysis and test maintenance for binary items*. Mooresville, IN: Scientific Software.
- Zwick, W.R., & Velicer, W.F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99*, 432-442.

Table 1
Content Specifications for SAIP Science 1999 assessment

	Item Ability Level (Test Item Difficulty)				
	1	2	3	4	5
Knowledge: Biology	<i>B44, B38,</i> B40, B1	B27, B41, B37, B26	A2, B28, C36, A1, <i>B7-C47</i>	<i>C37, C38,</i> C56, C16	C35, C57, C7, <i>C54</i>
Knowledge: Chemistry	B20, B55, B19, B6	<i>B54, B51,</i> B42, B15	<i>B34-C13, A7,</i> B59-C7, A8	<i>C11, C17,</i> C51, C34	C53, C52, C12, C50
Knowledge: Earth	B65, B36, <i>B28, B66</i>	<i>B48, B60,</i> B45, B47	B10-C1, A12, <i>B2-C22, A9</i>	<i>C6, C55, C4,</i> C18	<i>C63, C25,</i> C24, C41
Knowledge: Physics	B56, B49, B22, B21	B57, B23, <i>B25, B61</i>	<i>B58-C15, A5,</i> B18-C49, A3	C20, C9, C29, <i>C61</i>	<i>C60, C62,</i> C28, C10
Nature of Science	B43, B29, <i>B4, B39, B5</i>	<i>B32, B9, B12,</i> <i>B31, B33</i>	<i>B30-C44,</i> <i>A10, B63-C21,</i> <i>B11-C2, A11</i>	<i>C46, C33,</i> <i>C32, C30,</i> <i>C39</i>	<i>C65, C64, C5,</i> C66, C31
Science, Technology, and Society	B17, B46, <i>B7, B53, B13</i>	<i>B3, B16, B24,</i> B8, B50	B52-C45, A6, B35-C14, A4, <i>B62-C26</i>	C42, C27, <i>C28, C59, C3</i>	C23, C43, C19, C58, <i>C40</i>

Note. Bold font items are Conceptual items, regular font items are Procedural items, and italicized font items are Use items. (Hyphenated questions are duplicates)

Table 2
Sample Size for Eight Data Files by Section, Age and Cross Validation Group

	Section AB		Section AC	
	13 year-olds	16-year-olds	13-year-olds	16-year-olds
Group 1	3054 (AB1-13)	2000 (AB1-16)	2903 (AC1-13)	4010 (AC1-16)
Group 2	3054 (AB2-13)	2000 (AB2-16)	2903 (AC2-13)	4010 (AC2-16)

Note. Data file name in parenthesis

Table 3
DIMTEST-FAC Analysis Results

	Test Section							
	Section AB				Section AC			
	Sample size for FAC	Sample size for T	T	P	Sample size for FAC	Sample size for T	T	P
13-year-old group1	1000	2054	4.8891	0.0000	800	2103	4.1602	0.0000
13-year-old group2	1000	2054	5.4745	0.0000	800	2103	4.3294	0.0000
16-year-old group1	800	1200	3.5009	0.0002	1000	3010	4.2448	0.0000
16-year-old group2	800	1200	3.5425	0.0002	1000	3010	5.9118	0.0000

Table 4
The Number of Factors Retained By Section, Age and Group Using Five Decision Rules

Rule	Section AB				Section AC			
	13-year-olds		16-year-olds		13-year-olds		16-year-olds	
	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2
K1	24	25	27	27	30	30	27	27
Scree	2	2	2	2	2	2	1	1
Kaiser	3	3	4	4	2	2	2	3
PA	16	16	19	19	25	25	20	19
MAP	3	3	2	2	2	2	3	3

Table 5

Factor Loadings from Exploratory Factor Analysis of SAIP Science 99 Assessment after Oblimin Transformation, Section AB (Group 1) for 13-year-olds by Content Description and Ability Level

Item	Description	Code	Ability Level	Factor 1	Factor 2	Factor 3
4	ST	6	3	.406		
6	ST	6	3	.437		
13	CB	1	1	.345		
14	CE	3	3	.405		
15	ST	6	2	.312		
16	NS	5	1	.517		
18	CC	2	1	.516		
19	ST	6	1	.551		
20	ST	6	2	.532		
21	NS	5	2	.538		
22	CE	3	3	.346		
25	ST	6	1	.604		
26	CB	1	3	.657	-.308	
27	CC	2	2	.329	.657	
28	ST	6	2	.421		
29	ST	6	1	.447	.374	
30	CP	4	3		.307	
31	CC	2	1	.449		
32	CC	2	1	.489		
33	CP	4	1	.451		
34	CP	4	1	.337		
35	CP	4	2	.387	.318	
36	ST	6	2		.306	
37	CP	4	2	.437	.328	
38	CB	1	2		.595	
39	CB	1	2	.352		
40	CE	3	1	.317		
41	NS	5	1	.479		
42	NS	5	3	.301		
43	NS	5	2	.443		
45	NS	5	2	.536		
47	ST	6	3	.406	.384	
51	NS	5	1	.479		
52	CB	1	1	.451	-.370	
54	CC	2	2	.507		
55	NS	5	1	.622		
56	CB	1	1	.560		
57	CE	3	2	.486		
58	ST	6	1	.360		.490
59	CE	3	2	.379		
60	CE	3	2	.400		
63	CC	2	2	.317		
64	ST	6	3	.397		
65	ST	6	1	.427		
66	CC	2	2	.525		
67	CC	2	1	.388		.405
68	CP	4	1	.354		
70	CP	4	3			.516
71	CC	2	3	.377		-.541
72	CE	3	2	.323		
73	CP	4	2			.324
74	ST	6	3	.474		-.338
75	NS	5	3			.598

Table 5a

Factor Loadings from Exploratory Factor Analysis of SAIP Science 99 Assessment after Oblimin Transformation, Section AB (Group 2) for 13-year-olds by Content Description and Ability Level

Item	Description	Code	Ability Level	Factor 1	Factor 2	Factor 3
4	ST	6	3	.454		
6	ST	6	3	.458		
13	CB	1	1	.339		
14	CE	3	3	.393		
15	ST	6	2	.337		
16	NS	5	1	.464		
18	CC	2	1	.497		
19	ST	6	1	.626		
20	ST	6	2	.525		
21	NS	5	2	.569		
22	CE	3	3	.396		
25	ST	6	1	.597		
26	CB	1	3	.641	-.321	
27	CC	2	2	.355	.694	
28	ST	6	2	.421		
29	ST	6	1	.423	-.403	
30	CP	4	3		.332	
31	CC	2	1	.481		
32	CC	2	1	.513		
33	CP	4	1	.521		
34	CP	4	1	.352		
35	CP	4	2	.391	-.316	
36	ST	6	2		.329	
37	CP	4	2	.436	.357	
38	CB	1	2		-.619	
39	CB	1	2	.330		
41	NS	5	1	.489		
43	NS	5	2	.453		
45	NS	5	2	.545		
47	ST	6	3	.430		
51	NS	5	1	.487		
52	CB	1	1	.464	-.320	
54	CC	2	2	.506	.303	
55	NS	5	1	.676		
56	CB	1	1	.574		
57	CE	3	2	.541		
58	ST	6	1	.464		.392
59	CE	3	2	.384		
60	CE	3	2	.417		
63	CC	2	2	.377		
64	ST	6	3	.377		
65	ST	6	1	.416		
66	CC	2	2	.551		
67	CC	2	1	.436		.382
68	CP	4	1	.395		
69	CP	4	2	.320		
70	CP	4	3			.513
71	CC	2	3	.380		-.578
73	CP	4	2	.321		
74	ST	6	3	.429		-.400
75	NS	5	3			.599

Table 6

Factor Loadings from Exploratory Factor Analysis of SAIP Science 99 Assessment after Oblimin Transformation, Section AB (Group 1) for 16-year-olds by Content Description and Ability Level

Item	Description	Code	Ability Level	Factor 1	Factor 2
6	ST	6	3	.382	
7	CC	2	3	.311	
13	CB	1	1	.332	
16	NS	5	1	.366	
18	CC	2	1	.363	
19	ST	6	1	.561	
20	ST	6	2	.469	
21	NS	5	2	.446	
22	CE	3	3	.322	
24	NS	5	2	.307	
25	ST	6	1	.481	
26	CB	1	3		-.508
27	CC	2	2	.604	.314
29	ST	6	1		-.474
30	CP	4	3	.411	
31	CC	2	1	.372	-.302
32	CC	2	1	.543	
33	CP	4	1	.345	
35	CP	4	2		-.424
37	CP	4	2	.525	
38	CB	1	2		-.675
39	CB	1	2	.439	
41	NS	5	1	.466	
43	CP	4	2	.476	
45	NS	5	2	.414	
47	ST	6	3	.497	
51	NS	5	1	.459	
52	CB	1	1		-.319
54	CB	1	2	.565	
55	NS	5	1	.586	
56	CB	1	1	.431	
57	CE	3	2		-.407
58	ST	6	1		-.529
59	CE	3	2	.422	
60	CE	3	2	.407	
64	ST	6	3	.395	
67	CC	2	1		-.322
68	CP	4	1	.414	
69	CP	4	2		-.318
71	CC	2	3	.432	.301
72	CE	3	2		-.339
74	ST	6	3	.419	
75	NS	5	3		-.521

Table 6a

Factor Loadings from Exploratory Factor Analysis of SAIP Science 99 Assessment, after Oblimin Transformation, Section AB (Group 2) for 16-year-olds by Content Description and Ability Level

Item	Description	Code	Ability Level	Factor 1	Factor 2
6	ST	6	3	.381	
7	CC	2	3	.310	
13	CB	1	1	.332	
16	NS	5	1	.366	
18	CC	2	1	.363	
19	ST	6	1	.561	
20	ST	6	2	.469	
21	NS	5	2	.446	
22	CE	3	3	.320	
24	NS	5	2	.305	
25	ST	6	1	.479	
26	CB	1	3		-.509
27	CC	2	2	.604	.315
29	ST	6	1		-.475
30	CP	4	3	.410	
31	CC	2	1	.371	-.300
32	CC	2	1	.542	
33	CP	4	1	.344	
35	CP	4	2		-.425
37	CP	4	2	.524	
38	CB	1	2		-.676
39	CB	1	2	.441	
41	NS	5	1	.464	
43	NS	5	2	.475	
45	NS	5	2	.413	
47	ST	6	3	.499	
51	NS	5	1	.458	
52	CB	1	1		-.319
54	CC	2	2	.564	
55	NS	5	1	.586	
56	CB	1	1	.430	
57	CE	3	2		-.408
58	ST	6	1		-.530
59	CE	3	2	.421	
60	CE	3	2	.406	
64	ST	6	3	.398	
67	CC	2	1		-.323
68	CP	4	1	.413	
69	CP	4	2		-.313
71	CC	2	3	.431	.300
72	CE	3	2		-.340
74	ST	6	3	.418	
75	NS	5	3		-.522

Table 7

Factor Loadings from Exploratory Factor Analysis of SAIP Science 99 Assessment after Oblimin Transformation, Section AC (Group 1) for 13-year-olds by Content Description and Ability Level

Item	Description	Code	Ability Level	Factor 1	Factor 2
13	CE	3	3	.384	
16	CE	3	4	.356	
17	NS	5	5		.373
18	CE	3	4		.332
20	CC	2	3	.303	
26	ST	6	3	.330	
29	CC	2	4		.475
30	CE	3	4	.308	
31	ST	6	5		.425
33	NS	5	3	.455	
34	CE	3	3	.305	
35	ST	6	5	.407	
37	CE	3	5		.649
39	ST	6	4	.340	
42	NS	5	4	.435	
43	NS	5	5	.303	
44	NS	5	4	.333	
46	CB	1	5	.432	
47	CB	1	3	.309	
48	CB	1	4	.393	
49	CB	1	4	.313	
51	ST	6	5		.398
52	CE	3	5		.474
53	ST	6	4	.392	
54	ST	6	5	.454	
55	NS	5	3	.413	
56	ST	6	3	.453	
58	CB	1	3	.600	
59	ST	6	4	.308	.327
60	CP	4	3	.370	
65	CB	1	5	.331	
70	ST	6	4	.391	
75	NS	5	5		.470

Table 7a

Factor Loadings from Exploratory Factor Analysis of SAIP Science 99 Assessment after Oblimin Transformation, Section AC (Group 2) for 13-year-olds by Content Description and Ability Level

Item	Description	Code	Ability Level	Factor 1	Factor 2
13	CE	3	3	.451	
17	NS	5	5		-.435
18	CE	3	4		-.378
20	CC	2	3		-.311
29	CC	2	4		-.586
31	ST	6	5		-.421
33	NS	5	3	.338	
34	CE	3	3	.339	
35	ST	6	5		-.325
37	CE	3	5		-.606
42	NS	5	4	.421	
44	NS	5	4	.302	
46	CB	1	5	.374	
47	CB	1	3	.301	
48	CB	1	4	.399	
49	CB	1	4	.317	
50	NS	5	4	.303	
51	ST	6	5		-.314
52	CE	3	5		-.485
53	ST	6	4	.349	
54	ST	6	5	.388	
55	NS	5	3	.435	
56	ST	6	3	.418	
58	CB	1	3	.579	
59	ST	6	4		-.338
60	CP	4	3	.363	
65	CB	1	5	.331	
70	ST	6	4	.389	
75	NS	5	5		-.450
76	NS	5	5		-.383

Table 8

Factor Loadings from Exploratory Factor Analysis of SAIP Science 99 Assessment after Oblimin Transformation, Section AC (Group 1) for 16-year-olds by Content Description and Ability Level

Item	Description	Code	Ability Level	Factor 1	Factor 2	Factor 3
12	CE	3	3	.303		
15	ST	6	4			-.411
16	CE	3	4			-.363
17	NS	5	5			-.558
18	CE	3	4			-.381
20	CC	2	3			-.359
24	CC	2	5	.346		
26	ST	6	3			-.307
28	CB	1	4	.306		
30	CE	3	4	.304		
31	ST	5	5			-.420
32	CP	4	4	.357		
33	NS	5	3			-.333
35	ST	6	5			-.493
37	CE	3	5		.379	-.391
38	ST	6	3			-.330
41	CP	4	4	.319	.363	
42	NS	5	4	.381		
44	NS	5	4			-.369
47	CB	1	3	.495		
48	CB	1	4	.385		
49	CB	1	4	.337		
51	ST	6	5			-.343
52	CE	3	5			-.366
53	ST	6	4	.327		
54	ST	6	5			-.464
55	NS	5	3	.336		
56	ST	6	3	.397		
58	CB	1	3			-.303
59	ST	6	4			-.401
60	CP	4	3	.500		
62	CC	2	4	.466		
65	CB	1	5	.385		
66	CE	3	4	.430		
75	NS	5	5			-.489
76	NS	5	5			-.378

Table 8a

Factor Loadings from Exploratory Factor Analysis of SAIP Science 99 Assessment after Oblimin Transformation, Section AC (Group 2) for 16-year-olds by Content Description and Ability Level

Item	Description	Code	Ability Level	Factor 1	Factor 2	Factor 3
12	CE	3	3	.386		
15	ST	6	4			-.396
16	CE	3	4			-.466
17	NS	5	5			-.549
18	CE	3	4			-.454
20	CC	2	3			-.331
23	CC	2	4	.366		
24	CC	2	5	.399		
28	CB	1	4	.355		
29	CC	2	4	.379		
30	CE	3	4	.338		
31	ST	6	5			-.376
33	NS	5	3			-.331
35	ST	6	5			-.511
37	CE	3	5			-.460
38	ST	6	3			-.378
40	CP	4	5	.312		
41	CP	4	4	.339	.316	
42	NS	5	4	.335		
44	NS	5	4			-.382
47	CB	1	3	.503		
48	CB	1	4	.330		
51	ST	6	5			-.347
52	CE	3	5			-.362
54	ST	6	5			-.469
56	ST	6	3	.339		
58	CB	1	3			-.407
59	ST	6	4			-.368
60	CP	4	3	.500		
62	CC	2	4	.443		
65	CB	1	5	.353		
66	CE	3	4	.440		
75	NS	5	5			-.516
46	NS	5	5			-.409

Table 9
Confirmatory Factor Analysis Causal-Categorical 2-factor model

TEST	Number of items	Sample size		LISREL index		
		N	Valid N (listwise)	RMSEA	RMR	AGFI
AB 13-year-old group	52	6109	3666	0.057	0.055	0.79
AB 16-year-old group	43	3185	1992	0.074	0.066	0.75
AC 13-year-old group	33	5806	5758	0.051	0.044	0.90
AC 16-year-old group	36	8021	7942	0.047	0.042	0.91

Table 10
Confirmatory factor Analysis Item-Format 2-factor model

TEST	Number of items	Sample size		LISREL index		
		N	Valid N (listwise)	RMSEA	RMR	AGFI
AB 13-year-old group	52	6109	3666	0.057	0.055	0.79
AB 16-year-old group	43	3185	1992	0.073	0.066	0.76
AC 13-year-old group	33	5806	5758	0.048	0.041	0.92
AC 16-year-old group	36	8021	7942	0.042	0.038	0.93

Table 11
Confirmatory factor Analysis Test specification 6-factor model

TEST	Number of items	Sample size		LISREL index		
		N	Valid N (listwise)	RMSEA	RMR	AGFI
AB 13-year-old group	52	6109	3666	0.057	0.056	0.79
AB 16-year-old group	43	3185	1992	0.074	0.066	0.76
AC 13-year-old group	33	5806	5758	0.050	0.044	0.91
AC 16-year-old group	36	8021	7942	0.044	0.040	0.92