The Role of Rasch Analysis When Conducting Science Education Research Utilizing Multiple-Choice Tests

WILLIAM J. BOONE
School of Education and Allied Professions, Miami University, Oxford, OH 45056, USA

KATHRYN SCANTLEBURY
Department of Chemistry and Biochemistry, University of Delaware, Newark, DE 19716, USA

Received 29 June 2004; revised 28 May 2005; accepted 20 June 2005

DOI 10.1002/sce.20106
Published online 15 November 2005 in Wiley InterScience (www.interscience.wiley.com).

ABSTRACT: Recent international studies note that countries whose students perform well on international science assessments report the need to change science education. Some countries use assessments for diagnostic purposes to assist teachers in addressing their students' needs. However, in the United States, standards-based reform has focused the national discussion on documenting students' attainment of high educational standards. Students' science achievement is one of those standards, and in many states, "high-stakes" tests determine the resultant achievement measures. Policymakers and administrators use those tests to rank school performance, to prevent students' graduation, and to evaluate teachers. With science test measures used in different ways, statistical confidence in the measures' validity and reliability is essential. Using a science achievement test from one state's systemic reform project as an example, this paper discusses the strengths of the Rasch model as a psychometric tool and analysis technique, referring to person item maps, anchoring, differential item functioning, and person item fit. Furthermore, the paper proposes that science educators should carefully inspect the tools they use to measure and document changes in educational systems. *Sci Ed* **90:**253–269, 2006

INTRODUCTION

In the 1980s, policymakers in several Western countries began questioning the effectiveness of their public education systems (National Commission of Excellence in Education, 1983; Gregory & Clarke, 2003). In the United States, this discussion began with the

Correspondence to: William J. Boone; e-mail: boonewj@muohio.edu
Contract grant sponsor: National Science Foundation.
Contract grant number: REC 9602137.
The opinions expressed are those of the authors and do not necessarily reflect the position of NSF.

publication of *A Nation at Risk* that proposed a systemic approach to improving American education using standards-based school reform (Clewell et al., 1995; National Commission of Excellence in Education, 1983). In Britain, politicians have questioned the ability of Local Education Authorities (LEAs) to produce educated citizens because of wide variation in students' achievement and social conformity. The resultant Education Reform Act of 1988 produced a national curriculum and a student assessment program, from which school results were published (Gregory & Clarke, 2003). Atkin (1998) reported that regardless of how students' scored on the Third International Mathematics and Science Study (TIMSS), all of the countries that participated in the OCED's international study of science and mathematics curriculum reform were dissatisfied with their current programs. For example, Japan and Germany developed new curriculum to educate students on conservation issues. And for most countries in the study, developing curriculum that related science to everyday settings, integrating science with other subjects, and changing teachers' pedagogical strategies from an emphasis on lecturing to involving students in through hands-on activities were other key issues (Atkin, 1998). These reforms were also the foci for science education in the United States.

Over the past 50 years, three United States reform efforts in science education have focused on as texts and teaching, courses, and curriculum, and the most recent effort built upon and combined educators' knowledge and experience from the first two waves to develop a theoretical framework that focused on equity and excellence (Kahle, in press). For science education, a consequence of this dominating philosophy and approach to improving education was that the National Science Foundation (NSF) developed and funded a series of large projects directed toward systemically reforming K-12 mathematics and science. One focus of the NSF projects was the improvement of students' achievement in science and mathematics as well as decreasing a widening achievement gap between served and underserved students (NSF, 1998).

In the past two decades, student outcome data, such as achievement scores on state-mandated and/or standards-based tests have become "high-stakes" and an integral component of many countries and state's accountability procedures (Olson, 2001). In the United States, many high-stakes tests are mandated by states (e.g. Delaware's Delaware Student Testing Program, and/or federal law (No Child Left Behind (NCLB), Boone & Donnelly, in review). Due to the high cost of administering and scoring statewide assessments, the majority of such assessments are composed of multiple-choice test items (e.g. Indiana's Indiana Statewide Testing for Educational Progress (ISTEP) examination).

Further, studies have shown that numerous teacher factors, such as teachers' certification, content preparation, effectiveness, and their belief that underrepresented and underserved students are capable in science influenced students' achievement (Darling-Hammond, 2000). Through their involvement with teacher professional development programs, membership on state and national testing councils, and service within their local school districts, science educators are pivotal leaders in the reform effort. And with their research expertise and skills, science educators can assist teachers, administrators, and school district personnel to study reform efforts at a local level, to interpret how curriculum programs may change because of student achievement on state, national or international tests (e.g., TIMSS), or to suggest how standards, assessment, and curricula are aligned to achieve the dual goals of improving the achievement of all students and closing the gaps that exist between subgroups of students.

Part of a science educator's research expertise is developing, selecting, and using the appropriate measurement tools to ascertain if educational systems are improving based upon reform criteria. In order to achieve this goal, researchers need robust, yet sensitive, psychometric tools that can measure whether reform efforts are producing the necessary changes

that will positively influence student achievement. This paper discusses the advantages and limitations of using the Rasch measurement model to study large-scale science education reform by providing an overview of the psychometric steps that influenced the statewide collection and analysis of a science achievement test. In particular, we will discuss measurement issues for underserved groups in science, such as females and African Americans. (For a discussion of item selection, content and format, see Kahle, 1997; Kahle, in press; Kahle, Meece, & Scantlebury, 2000; Scantlebury et al., 2001.) Using data collected for *Discovery*,[1] we will discuss the psychometric measurement issues confronted by the project and the critical importance of those issues with respect to evaluating student achievement data. Rasch measurement provides researchers with statistical tools to evaluate different and diverse data sets. In this paper we discuss one kind of data set, (the multiple-choice test), because this data type is commonly used for evaluating large student samples.

BACKGROUND

Since the development of Rasch's measurement model (Rasch, 1960) and its subsequent expansion by Wright and others (e.g., Rentz & Bashaw, 1977), many investigators have utilized the model to develop tests and to calculate respondent measures. The Rasch model provides valuable data for the development, modification, and monitoring of valid measurement instruments for industry, medicine, and educational research communities. Further, the Rasch model provides equal interval scales, which are invariant. In this context, invariant means that the scale defined by items can be monitored so that the items define the latent trait in the same manner from time point to time point. This allows researchers to confidently compare results over time. Many medical groups use this model for board certification and recertification tests (e.g., American Dental Association, American Board of Pediatric Dentistry, American Society of Clinical Pathologists, American Board of Medical Examiners), while Rehfeldt's (1990) work provides an industrial example (evaluation of paint products), there are numerous national and international educational examples. In Australia, the model has been widely used to evaluate student performance and communicate results to parents (e.g. Australian Council of Educational Research, 1995). In recent years, large-scale assessment projects such as TIMSS, and the long-term evaluation of reform in the Chicago Public Schools (Bryk et al., 1998), have used the Rasch model. In the latter ongoing study, Bryk et al. (1998) conducted Rasch calibrations prior to utilizing parametric tests and Hierarchical Linear Model (HLM) analysis. Kahle et al. (2000) also used Rasch measures when analyzing the effect of teachers "professional development experiences on African American students" science attitudes and achievement.

MEASUREMENT ISSUES IN SCIENCE EDUCATION

In science education, test development and utilization usually includes the following steps: (1) experts write, edit, and/or select potential items; (2) test reliability is determined through the calculation of a single statistic (e.g., Cronbach alpha); (3) student performance is calculated using raw scores; and (4) parametric tests such as *t*-tests and ANOVAs are conducted using student raw scores. Often published science education research provides minimal information regarding item construction, item selection, and test development. Authors rarely discuss the validity of using respondents' raw scores, the role of measurement error, and the rationale for longitudinal monitoring procedures. Commonly used Rasch

[1] *Discovery* was funded in 1991 by the National Science Foundation as Ohio's Statewide Systemic Initiative (Kahle, 1999). The state continued to fund the project after federal support ended in 1998.

modeling programs provide a raw score to equal interval measure table as well as a plot (curve) that show the nonlinear relationship between the two scales. The curve (o-give) describing the relationship demonstrates the nonlinear relationship between the two scales (one scale is interval and one is not).

In this paper, we will discuss the issues associated with using raw scores and provide an overview of how the Rasch model improves the measurement function of a multiple-choice science test. The Rasch model also determines the item difficulty independent of test takers, and the student's ability independent of test items. These characteristics of the model provide additional rigor to, and confidence in, students' computed scores (measures) on multiple-choice science tests.

How the Rasch Model Solves a Measurement Problem

Advances in science education will in part result when rigorous instruments are developed that provide useful equal interval measures that operate in a similar manner when used on different samples. To understand the strengths of a robust measurement instrument, it is helpful to consider the characteristics of a meter stick. Although the analogy is simple, it illustrates often forgotten, yet, critical measurement issues. The first strength is that a meter stick provides equal interval data. For example, a child's change in height from 90 to 100 cm represents exactly twice the change in height of another child whose height has changed from 75 to 80 cm. Regardless what part of the meter stick is used, the data collected with the meter stick are additive. A second strength is that the meter stick's accuracy is typically independent from its place of manufacture. That is, there is an international standard as to the ordering and spacing of centimeter marks. This agreed upon standard insures that data collected from a meter stick produced are one place combined with data collected with a meter stick produced at a different location. These two meter sticks may not be identical in that one may have marks every centimeter, while the other may have marks every half of a centimeter, but key marks are at the same location, thus allowing researchers to combine the measurements made with the two devices into a single data set.

For effective science education reform, it is crucial to use data expressed on an equal interval (linear) scale. Instruments used to measure the progress of educational reform, especially student achievement, need to be carefully developed and continually monitored. Perhaps most significantly, instruments must not measure different subgroups of students in different ways. It is the Rasch model which allowed Ohio's SSI to confront this issue. In the next section, we will explain how the Rasch model can create and monitor a "meter stick" defined by science test items.

The Rasch Model

The Rasch model developed by the Danish mathematician George Rasch (Rasch, 1960) differs from many other statistical models because it is a probabilistic model. The model explains how a person's performance with regard to a specific trait can predict that person's response (e.g., right or wrong) on a particular test item involving that trait. The issue that is being considered, be it science learning, science inquiry, or attitudes toward science, is often called the "latent trait." The mathematical model that describes this relationship between items and person is expressed by

$$\log[p_{ni}/(1 - p_{ni})] = B_n - D_i \tag{1}$$

where B_n is the ability of a person n and D_i is the difficulty of item i. As can been seen from Equation (1), the probability (p) of a person correctly answering an item is solely

related to her/his ability (with respect to the latent trait) and the difficulty of the item being answered. The model fulfills Thurstone's (1928) requirement for scale validity. Wright and Stone (1979) and Wright and Masters (1982) provided specific details regarding the model and through the writing of researchers such as Andrich (1978).

Figure 1 presents the logistic o-give curve that expresses the relationship between test takers' raw scores and the logit measures. Logit measures are the conversion of raw scores to logits through use of the Rasch model (Equation (1)). This figure illustrates the relationship between Rasch measures and raw scores. Using Figure 1, we can compare a student who has scored a 26 on a 28-item test to one who has scored 27 on that same test. In terms of raw score, those two students differ by 1 point, and they differ by approximately 1.0 logits (the units used in Rasch measurement to express ability). If we now compare two students, one with a score of 15 and the other with a score of 16, we can see that their raw scores also differed by 1, but their logit scores differ by 0.10. Logits calculated from the Rasch model allow researchers to avoid using nonequal interval values in parametric analyses that assume linearity.

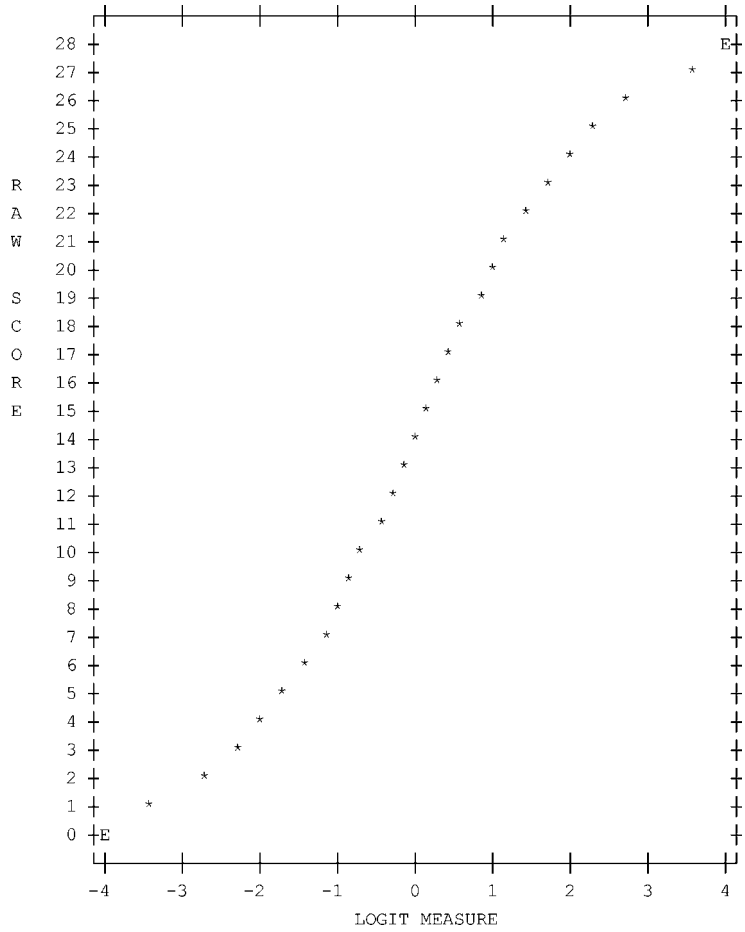

Figure 1. The relationship between raw scores and logit measures for the Year 1 science test. Note the nonliner relationship between raw scores and logits. Also note that differences in raw score totals do not equal the same difference in logit measures. Thus, before parametric tests are utilized, raw scores must be converted into logit measures.

The conversion of raw scores to equal interval measures is particularly important because many science reform efforts focus on monitoring the performance of underachieving and underrepresented students (Lee, 2002; U.S. Department of Education, 2004). If researchers do not convert raw scores to equal interval measures then the results of their analysis may provide incorrect and/or incomplete information on student performance. Figure 1 shows the nonlinear (noninterval) relationship of raw scores at the higher and lower ends of the ability scale. Many underrepresented groups in science fall at the low end of the science scales used to provide measures to researchers. As Figure 1 shows, it is at the very low and very high end of the raw score scale that the nonlinear relationship between raw scores and equal interval measures is the most pronounced. If a researchers uses only raw scores, then incorrect conclusions may be reached by using raw score data for parametric tests of student groups (for instance, poorly performing males and higher performing females).

Person–Item Maps

The Rasch model transforms raw item difficulties and raw person scores to equal interval measures. These measures are used to map persons and items onto a linear (interval) scale. Such mapping (called person–item maps) produces useful tools for evaluating reform and allows researchers to evaluate the instrument's effectiveness. Figure 2 is the person–item map from the Rasch analysis of the Year 1 Ohio SSI science test. This analysis provides equal interval measures in logit units of student performance and item difficulty. Items ranged from easiest (base of the graph) to hardest (top of the graph). Persons are plotted as a function of their ability, with the more able students at the top of the figure, and less able students at the base. A student plotted at the same logit level as an item is a person who has an ability level at the difficulty level of the item. Based on the Rasch model, that student has a 50/50 chance of correctly answering that item. Items plotted above any person are harder than the person's ability level. For those items, there is less than a 50% chance the student will correctly answer any items above their ability level. Similarly, items below a person are those items for which the person has a greater than 50/50 chance of correctly answering.

The person–item map provides a wealth of information for building and monitoring a test that is useful in an analysis of science education reform. First, the item component of the person–item map helps in the identification of items that involve the same portion of the latent trait (the same portion of the variable being measured by the test). To return to our meter stick example, cutting a mark into the meter stick two or three times in the identical location does not improve the meter stick's measurement precision. Test developers can remove redundant items and maintain the test's integrity. The same is true when test items involve the same portion of a single latent trait. In the *Discovery* study, after analyzing data from the Year 1 science test, the researchers shortened the Year 2 science test to minimize the administrative burden on teachers and loss of instructional time. Researchers focused on items of the same difficulty level and used the item overlap shown in Figure 2, as a guide in choosing which test items to remove.

A consideration of equity issues also led researchers to shorten the test. Boone (1998) conducted an analysis of test-taking patterns as a function of race and gender on the Year 1, 28-item science test. A significantly larger number of females and African Americans did not answer items at the end of the test, compared with males and White students. Since the test's purpose was to document students' science achievement rather than their test-taking patterns, *Discovery's* research group shortened the test and did not grade omitted items as wrong. Students may not complete a test for many reasons, such as reading level, familiarity with item content, and motivation. Understanding and identifying students' different

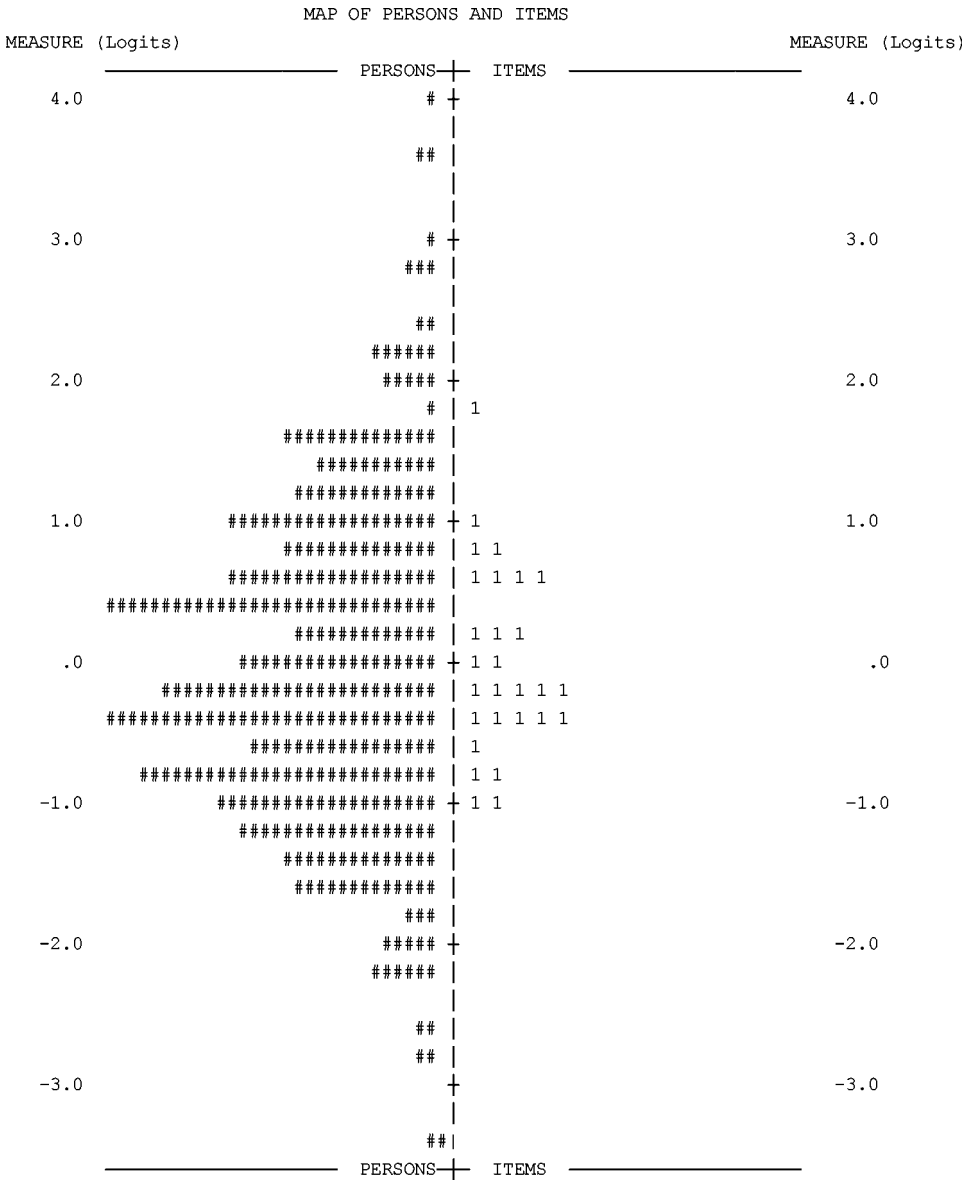

Figure 2. Persons measured with the Year 1 test and test items displayed on the same logit scale. Persons are represented by the # sign. Each of the test's 28 items is represented by the number "1." Persons at the same level as an item have a 50% chance of correctly answering that item. Items above their ability level can still be answered correctly, but students have less than a 50% chance of correctly answering the item. Items listed below a student are those that the student has less than a 50% chance of correctly answering. Each # sign represents two students.

test-taking patterns following a review of data is important; however, the key point is that by using Rasch techniques *Discovery* did not penalize those students who did not answer a subset of test items.

The person–item maps allowed researchers to make an informed decision about removing test items without jeopardizing the test's integrity. Figure 3 shows the distribution of items for the Year 2 science test. Compared to the Year 1 test, researchers removed test items that sampled the same part of the latent trait. For example, the Year 1 test had a large number

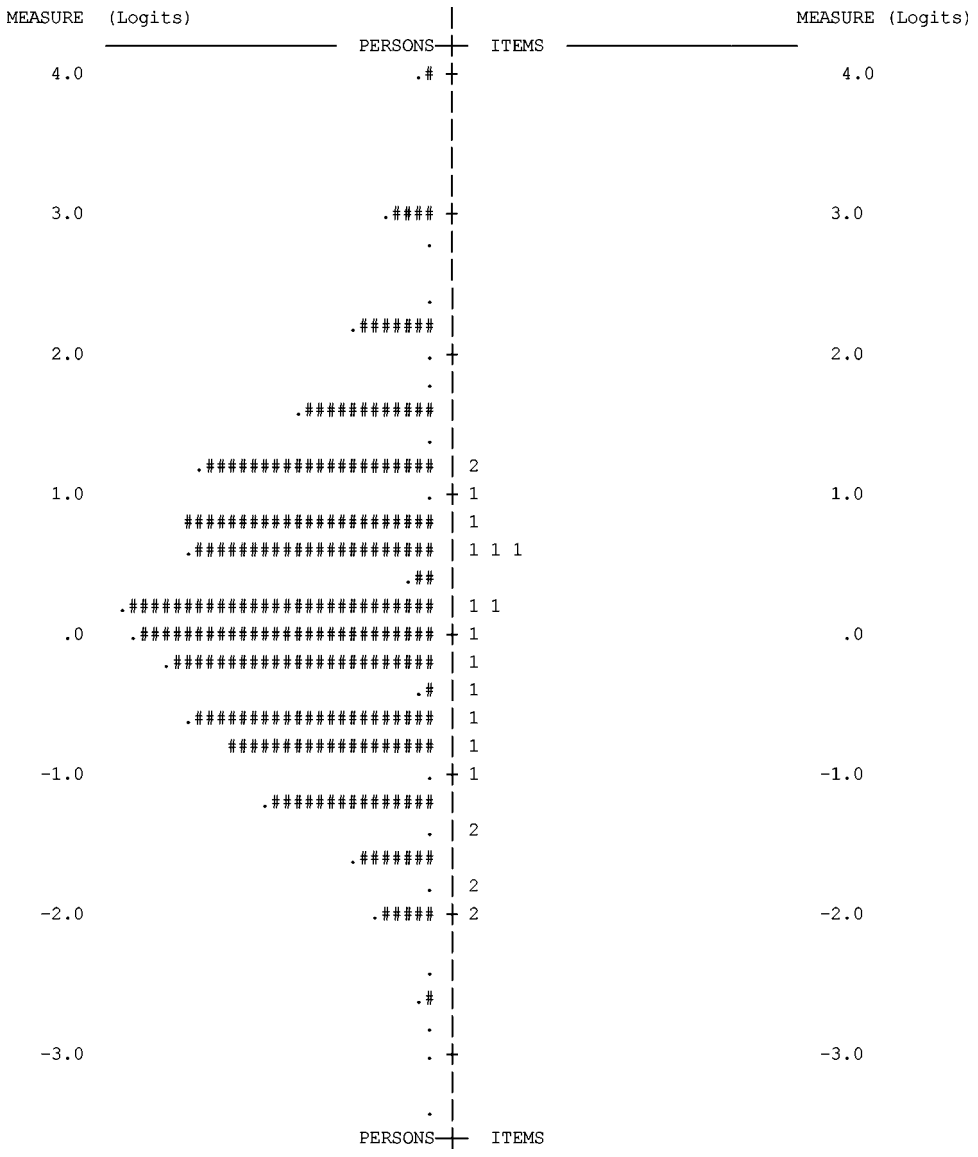

Figure 3. Persons measured with the Year 2 test and test items displayed on the same logit scale. Persons are represented by the # sign. Items labeled with the number 1 are items which were presented in the Year 1 and Year 2 test. Items which only first appeared in Year 2 are presented with the number 2. Persons at the same level as an item have a 50% chance of correctly answering that item. Items above their ability level can still be answered correctly, but students have less than a 50% chance of correctly answering the item. Items listed below a student are those that the student has less than a 50% chance of correctly answering. Each # sign represents a 11 students, while each "." represents 1 – 10 students. Notice that overall a greater range of items in terms of difficulty is presented to students in Year 2. Also students completing the test take fewer items which define the same portion of the latent trait (there are fewer items piled up at one difficulty level). Since common items are presented in Year 1 and Year 2, student performance in Year 2 can be expressed on the same equal interval logit scale as those students in Year 1. The gaps between columns of persons in the plot of Year 2 test takers result from the vertical scale selected, and from the fact that there were a much smaller combination of measures possible for Year 2 students due to the smaller number of items presented on the shortened test.

of items with a difficulty level between 0.0 and −0.5 logits. By removing several of those items from the Year 2 test, the test measured the same expanse of the latent trait, but with fewer items.

The ordering and spacing of items provides additional guidance for instrument developers. It is not only important to evaluate sections of the latent trait in which science items overlap but also regions that are devoid of items (see Figures 2 and 3) must be evaluated. These gaps represent regions in which adding science items would improve the instrument's measurement of students. For the Year 1 data, the gaps are notably at both ends of the defined latent trait. These gaps indicated that researchers could achieve a more balanced test by the addition of easy and hard items. Adding such items improves the possibility that the test would help differentiate the low and high achieving students from their peers. In Year 2, researchers added items to fill the gaps shown in Figure 2 and further differentiated students. Figure 3 shows the distribution of the Year 1 items on the Year 2 test, as well as the distribution of new, Year 2 items. Overall, many gaps in the latent trait present in Year 1 were filled in Year 2.

Figures 2 and 3 for the Year 1 test and the Year 2 test illustrate the importance of graphically presenting the spacing of both items and persons. Often, researchers do not present quantitative data in a form that facilitates evaluation. Tabular representation of persons or items provides an ordering, but it is difficult to qualitatively visualize the spacing between items and/or persons. Furthermore, the ordering and spacing of items can tell many stories for example, topic difficulty.

The person–item science plot provided *Discovery* researchers with multiple ways to evaluate and interpret the data. The ordering of items illustrates which concepts are easier or more difficult for students. Item spacing describes the range of students' understanding along the measured latent trait. For example, does the ordering and spacing of the items and people make sense? How, if at all, does the ordering and spacing of items deviate from predicted patterns? In the analysis, the researchers may use these maps to highlight possible problems in the test and/or data set. For instance, an unexpected ordering of science items may reveal miscoding, an incorrect key, or a misappraisal of student ability level. Presenting test item data in a simple composite plot with student performance guides researchers through the reappraisal of a test from initial development through field testing and re-administration.

The above steps provide researchers with a quick methodology for improving an instrument over time. The improvement of item targeting provided *Discovery* better differentiation between students and minimized the measurement error of student placement along the latent trait. Often simple, quick, improvements in the presentation of test items can allow researchers to uncover hidden data trends.

Anchoring Tests, Equating Tests

The Rasch model facilitates useful appraisals of person and item distribution and facilitating an assessment of the interaction of persons and items. Further, the Rasch model allows item removal and/or addition while retaining the same metric for comparing students across different cohorts and years by using item anchoring. After Year 1 *Discovery's* researchers removed selected items to shorten the science test. However, item anchoring allowed researchers to use the same metric, to compare student cohorts across the years. Item anchoring is "setting" items to a particular location along the latent trait. In the Year 2 test, each item's logit value was determined from the Year 1 data. This process insured that the metric used in Year 1 and Year 2 was the same. This flexibility is an important reason why researchers should consider the Rasch model for science instrument development. Although the Year 2 science test had fewer items compared with the Year 1 test, Rasch

anchoring allowed researchers to compare students' achievement between the two years. Thus, students who earned a particular raw score on one test are expressed on the same metric as those students who completed a revised test. There is no advantage or disadvantage if an individual takes a harder or easier test. This measurement technique allows researchers to refine the instruments throughout a reform effort, while retaining the same metric. The ability to constantly revise a measurement instrument improves its reliability and validity.

Anchoring with repeating items allows comparison with the same metric. We will use the meter stick analogy to illustrate anchoring. A 1-m stick can measure the height of students in Year 1 and a different meter stick used in Year 2. The two, meter sticks can differ in the number and/or distribution of marks. However, some tick marks representing a common length must be present on both meter sticks and must be the same distance from the origin of each meter stick. Thus if a tick mark common on one meter stick is for 5 cm, that tick mark must be 5 cm from the origin on both meter sticks. Furthermore, some common tick marks on both meter sticks must be present.

Discovery's researchers used anchoring technique in the following manner for the science test. In Year 2, researchers added easy and hard test items to improve the measurement of the latent trait (science ability). Although new, researchers used these items to calculate the Year 2 students' ability level and compare those students with the Year 1 cohort. In terms of the meter stick analogy, the Year 1 yardstick had a set number of tick marks and the Year 2 meter stick had fewer tick marks. Some Year 2 tick marks were in the same location (items had been retained from Year 1 to Year 2), while others were new (items had been added that would allow one to more clearly differentiate students). The common tick marks allow us to make comparisons between the Year 1 and Year 2 student cohorts. Further monitoring of the Year 2 test showed that there was no need to add or delete items. Figure 4 provides a schematic of linking tests from one year to another of a multiyear project.

Figure 4 summarizes linking test items and items that reappear from year to year. These items serve as a link throughout the time that *Discovery* has collected student achievement. However, another link uses items that are present only on adjacent years or those present only for a subset of years. These links help researchers develop science tests on the same metric, while providing the flexibility that allowed the inclusion or removal of items from year to year.

In the case of *Discovery's* efforts to evaluate reform across multiple years through common items in Year 1 and Year 2, this linking technique allowed researchers to place all test takers on the same scale. Researchers could compare students who took the longer 28-item test in Year 1 to those who took the shorter 18-item test over subsequent years. After

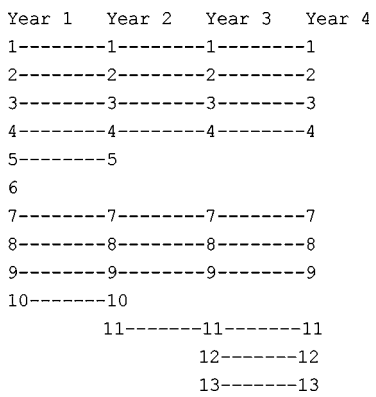

Figure 4. An example of item linkings from year to year. Numbers represent test items.

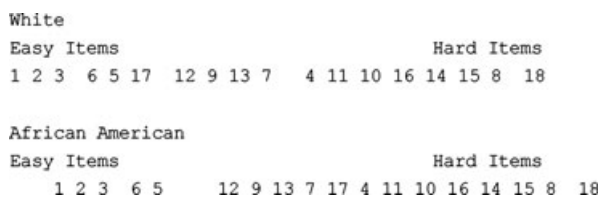

Figure 5. The distribution of items on the Year 2 test as a function of racial group. One item, 17, exhibits a statistically different location with respect to the other test items. Remaining items, for both groups, define the latent trait in a similar manner. This can be seen in the similar spacing and ordering of items from easy to hard.

evaluating and revising the initial science test, data from one version of the science test measured students on the same scale as subsequent versions. This appraisal provides flexibility in test administration over time while maintaining and/or improving the test reliability and validity.

Differential Item Functioning

A key issue in the development of any instrument is to identify biased test items, such as items that are easier for one group (e.g., males) to answer. Typically, experts provide an assessment of face validity regarding the test item bias. However, Differential Item Functioning (DIF) is also a second review for test item bias. Why investigate DIF? The consistent ordering and spacing of items that measure the latent trait for subgroups of individuals provides researchers with useful information to evaluate the instrument. Figure 5 presents a DIF analysis on the Year 1 science test for African American and White students. In order to investigate DIF for subgroups, researchers used the Rasch model to separately evaluate the science test items for African American and White students. The separate item calibrations for the two subgroups were plotted against each other. If an item's location with respect to the other test items within the subgroup deviated statistically from the pattern observed for the other subgroups, then that item was viewed as possibly defining a statistically different part of the latent trait for one group in comparison to the other group. When researchers identified item as possibly exhibiting DIF, they reviewed the item for bias such as item wording. In some cases, an item with possible DIF was removed before computing students' mean scores. When there was insufficient evidence of consistent bias to warrant removal of the item from the test, researchers' monitored the item for subgroup of students.

Typically, researchers do not monitor nonbiased items. Researchers should calculate DIF for each administration of the measurement tool to identify item bias. Figure 5 displays schematically the result of considering the distribution and spacing of test items determined by evaluating only White students and only African Americans taking one of the *Discovery* science tests. The figure shows that although the vast majority of items were slightly more difficult for African American students, the ordering and spacing of items for the two groups is about the same. However, for one item, #17,[2] there was a statistical difference between where #17 is located with regard to the other test items when considering the two subgroups of students.

It is important to note that another reason for observing DIF may be related to the meaning of items changing over a period of time (that is, item drift). This change of meaning may

[2] The *Discovery* science test is secured because it is currently use to continue monitoring the long-term impact of Ohio's reform efforts in science.

result from curricular changes, implementation of state reforms, or the release of items in the public domain. For example, as participants in a teacher education program learn more about "inquiry," the meaning of an item that measures their attitudes toward inquiry might change. Thus, researchers should monitor the appropriateness of items changing meaning throughout the life of an instrument. In the *Discovery* science test, yearly monitoring suggested that one item appeared susceptible to item drift. Researchers retained the item on the test, but it was allowed to "float." This meant that the item was not set at a specific calibration for all 4 years of testing. As long as one has a range of items that are common from year to year, generally some items can be allowed to float without affecting to the researchers' confidence in comparing student measures from different cohorts.

Traditional Statistics

Rasch statistics provide similar psychometric information to traditional analyses. A point biserial expresses item discrimination, and a "person separation index" provides a statistic similar to the commonly used KR-20 reliability test. There are, however, a number of key problems with the KR-20 used in the science education literature. First, researchers use nonequal interval raw scores to calculate the KR-20. Second, extreme scores often are included in calculations of reliability, but since extreme scores have no score error variance, the effect increases reported reliability.

Classical test theory provides a single standard error of measurement (SEM). However, in Rasch measurement each item and test taker is provided an error term. If either a very large or small percentage of students correctly answer items, this produces larger errors than items targeted at the average ability level of the students. One can intuitively understand this difference in measurement error. When few people miss easy items, one cannot ascertain how easy the item is. Thus, easy items have very high levels of measurement error. The same is true of hard items. The researchers revising *Discovery's* science test considered the error of each item and the range of each person's error before item removal. The consideration of item and person errors reminded *Discovery* researchers that the measure of a person always has error and one must consider whether the error is small enough to facilitate the development of meaningful conclusions. In addition to person separation index, Rasch modeling also provides an item separation index. When the distribution of test items changes over time, this index can monitor any gains.

Person and Item Fit

One technique utilized in Rasch measurement is an evaluation of an individual person's responses to test items to the model known as "fit" statistics. For the *Discovery* science test, researchers calculated a single "fit" statistic for each student. This statistic evaluated how unexpected nature of the student's correct and incorrect answers. The fit statistic quantitatively reveals what, in traditional analysis, often requires an individualized, qualitative review of each student's answers. Figure 6 presents the items correctly answered by one student on the Year 2 test who had a "low fit statistic." The items are ordered from easiest

Easy Items										Hard Items							
1	2	3	6	5	12	9	13	7		4	11	10	16	14	15	8	18
C	C	C	C	C	C	C	W	W		W	W	W	W	W	W	W	W

Figure 6. The answers (correct-C, wrong-W) for a student completing the 17-item test. No items are unexpectedly answered correctly or incorrectly when the difficulty of items is compared to the answers of the student.

Easy																	Hard
Items																	Items
1	2	3	6	5	12	9	13	7	4	11	10	16	14	15	8		18
C	C	W	C	W	C	C	C	C	C	C	C	C	C	W	W		W

Figure 7. The answers (correct-C, wrong-W) for a student completing the 17-item test. Two items (3 and 5) are unexpectedly missed when the difficulty of items is compared to the answers of the student.

to hardest (based upon all students who completed the test). A review of this test-taker's completion of items of increasing difficulty, showed s/he did not unexpectedly miss items nor unexpectedly correctly answer items. Figure 7 presents a different student, one who had a "high fit statistic" because the student unexpectedly missed items 3 and 5 (two of the easiest items on the test), but correctly answered all the remaining items. Fit statistics provide a technique of quickly highlighting students who are, for whatever reason, idiosyncratic in their behavior. However, this statistic does not answer why students' answered items in an unexpected way, but it does suggest further investigation of particular respondents. Fit statistics are an important technique for identifying patterns in student answers often hidden in the data set. For example, does a particular type of student unexpectedly miss items? In the *Discovery* data set no clear patterns emerged.

In addition to highlighting unusual patterns in students' answers, fit statistics also allow the researcher to investigate item functioning. Point biserials are reviewed in traditional test development. However, after authoring test items, evaluation of items often does not continue. Item fit statistics (similar to person fit statistics) evaluate the predictability of test takers' answers, given their overall ability. Figure 8 shows a very small portion of the plot used to investigate the answers of 800+ students who answered item 5 on the Year 4 test.

Figure 8 shows that this item was quite easy for both low and high performing students. However, a significant number of able students missed the item. This result suggests that test developers should investigate the item's structure. As was the case for the investigation of person fit statistics, this statistic does not tell why a person unexpectedly answered (either correct or incorrect) the item, but it does highlight items that may need revision and/or may not define the latent trait. DIF detects nuances in items as well, and the use of person and item fit statistics as well as DIF can greatly strengthen the validity and reliability of a science test and provide researchers' direction on ways to improve and to monitor the instrument from one test administration to the next.

Fit statistics, as well as other aspects of the Rasch model, illustrate the importance of monitoring science test items and individuals over time, and using qualitative techniques to continue an investigation of persons and items highlighted by statistical tests. In particular, Rasch techniques, such as fit analysis, allow researchers to quickly and to efficiently review the performance of specific students and items.

Evaluating item and person fit also allows researchers to evaluate the extent that the data fit the Rasch model. The Rasch model assumes that there is a single trait (variable). If the data do not fit the model, for instance items and/or test takers misfit, then it does not make

Low Ability Students								High Ability Students		
Student #										
70	106	801	4	7	----------	99	102	66	71	22
c	c	c	c	c		c	w	w	w	c

Figure 8. Portion of a plot used to investigate student answers to one item on the Year 1 test. Due to space limitations only the five highest and five lowest ability students (based upon the test items) are presented.

measurement sense to pool all the items and/or all the test takers for a single analysis of data.[3]

Some researchers consider the Rasch model to be part of the family of Item Response Theory (IRT) models, and they argue that the Rasch model is a one-parameter (1P) model, and that other IRT models (2P and 3P) are equally helpful for the data analysis.

Are there limitations to a Rasch analysis of data? If a test is poorly designed, computing an overall measure using all test items may be impossible and results may only be evaluated at the item level. A basic Rasch analysis is not difficult, and the model provides researchers' many techniques to evaluate test reliability and validity. A Rasch analysis may take longer than a traditional analysis, but it provides a deeper understanding of instrument's strengths and weaknesses. A key benefit in using Rasch analysis with test development is the flexibility the technique provides for improving measurement precision. Researchers should carefully track items' appearance, revision, and removal from tests. A critical step was the development of coding schemes that correctly and succinctly identify items over time and through edited versions. Are there times when an analysis of raw data results in similar results as that determined from a Rasch data analysis? Figure 1 shows the relationship between the raw score on the *Discovery* tests and the equal interval Rasch measures. Those students earning raw scores in the middle of the test are located along the linear portion of Figure 1, and for those students using raw scores should not impact an analysis. It is however, at the very high and low end of the test, where the relationship is nonlinear that using raw scores would result in an incorrect evaluation of students. In high-stakes tests, policymakers and educators focus on the low-performing students. *Discovery* focused on providing standards-based professional development to middle school science teachers as a strategy for improving student achievement, thus it was critical to use valid measures that would for identifying high and low performing students. Further, equity and excellence were major goals for *Discovery,* thus it was important to use measures which could provide reliable student outcome data for different subgroups.

Where to Start? How to Start?

It is not difficult to initiate a Rasch analysis. The authors of this study currently utilize the user-friendly windows-based software Winsteps (Linacre, 2005). This software is inexpensive, and the author of that software provides frequent introductory workshops throughout the world. There are other Rasch analysis computer programs also available.[4]

SUMMARY

> Because the ultimate goal is clear, a system of indicators of . . . student achievement capable of measuring longitudinal change is a virtually indispensable element of systemic reform. (Clune & Webb, 1997, p. 7)

Science educators are involved in the current reform efforts aimed at improving the quality of teaching, increasing student achievement, and eliminating achievement gaps between subgroups of students. There are harsh ramifications for students and sobering

[3] Others have discussed the issue of data to model fit, and readers can refer to those references for additional guidance (Bond & Fox, 2001).

[4] If science education researchers are interested in conducting an analysis similar to that carried out for Ohio's State Systemic Initiative, contact the authors of this paper, or start by contacting the authors of Rasch software programs.

implications for policymakers if "high-stakes" testing data are of poor quality. For example, the New York City School District announced that thousands of students had unnecessarily attended summer school because of errors in their test scores (Goodnough, 2001). At an individual level, many states require students to attain minimal proficiency levels on state tests for high school graduation. In some states, achievement on a high-stakes test is the sole criteria for judging if students have attained the state proficiency levels (Education Week, 2001). However, research has shown that a researcher's analysis decisions could further disadvantage underserved students (Boone, 1998). We suggest that by using the Rasch model, science educators can improve the quality of quantitative measurement at the individual and the systemic level.

The Rasch model helps test developers confront the important issue of monitoring scales and allows researchers to change those scales over time. Anchoring the scales using selected common items means that with a modified measurement instrument researchers can still compare cohorts of students over time. This also means that measurement instruments may track the changes in student achievement based upon curricular changes and/or teacher professional development aligned with district and state standards. Second, the Rasch model helps researchers quickly identify and evaluate specific student subgroups and individuals. For example, we previously described how an item functioned differently for African American students compared with their White peers. Another advantage of Rasch analysis is researchers can "flag" students with unexpected performance on parts of the test. Teachers (and/or researchers) could interview those students regarding the science concepts covered on the test items that they had given unexpectedly answers. In this regard, the model is diagnostic and supports qualitative data techniques because it can identify specific students and schools for further study.

In many large-scale projects, researchers often collect more data than can be expected to be evaluated. Rasch statistics helped *Discovery's* researchers target data efforts and thus maximize time efficiency. The Rasch model is easy to use. Rasch "measures" of test-takers used for parametric tests, for those Rasch measures are equal interval metrics. Rasch measures also can be computed for attitudinal scales. There have been some examples of the use of Rasch measures to investigate science education problems (Keeves & Alagumalia, 1998; Sadler, 1999). It is important to note that all the techniques outlined for multiple-choice tests can be utilized with other types of common data collection devices. For instance, attitudinal surveys can be Rasch calibrated before respondent measures are used for parametric tests, along with the quality control methods outlined in this paper. DeSouza, Boone, and Yilmaz (2004) discuss how researchers used Rasch measures for the analysis of an attitudinal scale in the field of science education. The analysis of attitudinal data is more complex than the data evaluated for this paper, but the same basic steps are used.

Achieving educational equity, while maintaining high-academic standards, is a key objective of the current reform in science education. Using fit statistics, evaluation of person–item maps, and the study of DIF, the Rasch model provides researchers a number of techniques to investigate the functioning of scales with regard to subgroups of students. The anchoring of tests insures that comparisons of student achievement are based on the same metric over time. The Rasch model is a tool that allows researchers to investigate equity issues and minimize measurement bias. The model's strengths help science educators confidently conduct research.

REFERENCES

Andrich, D. (1978). Scaling attitude items constructed and scored in the Likert tradition. Educational and Psychological Measurement, 38, 665–680.

Atkin, (1998). The OECD study of innovations in science, mathematics and technology education. Journal of Curriculum Studies, 30, 6, 647–666.

Australian Council for Educational Research. (1995). Learning project assessment. Author: Melbourne, Victoria, Australia.

Bond, T., & Fox, C. (2001). Applying the Rasch model: Fundamental measurement in the human sciences. Mahwah, NJ: Lawrence Erlbaum Associates.

Boone, W. (1998). Assumptions, cautions, and solutions in the use of omitted test data to evaluate the achievement of under-represented groups in science-implications for long-term evaluation. Journal of Women and Minorities in Science and Engineering,4, 183–194.

Boone, W., & Donnelly, L. (in review). High stakes science testing. In K. Tobin (Ed.), An encyclopedia of science education. Westport, CT: Greenwood Press.

Byrk, A., Thum, Y., Easton, J., & Luppescu, S. (1998). Academic productivity of Chicago public elementary schools. Technical Report. Chicago, IL: Consortium on Chicago School Research.

Clewell, B., Hannaway, J., Cosentino de Cohen, C., Merryman, A., Mitchell, A., & O'Brian, J. (1995). Systemic reform in mathematics and science education: An urban perspective. Washington, DC: Urban Institute.

Clune, W., & Webb, N. (1997). An introduction to the papers and think piece themes. In W. Clune, S. Millar, S. Raizen, N. Webb, D. Bowcock, E. Britton, R. Gunter, & R. Mesquita (Eds.), Workshop Report No. 4, Research on systemic reform: What have we learned? What do we need to know? Synthesis of Second Annual NISE Forum Volume 1: Analysis (pp. 6–12). Madison, WI: National Institute for Science Education, University of Wisconsin-Madison.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. Education Policy Analysis Archives, 8. Retrieved on July 8, 2003 from http://epaa.asu.edu/epaa/v8n1.

Deboer, G. (2002). Student-centered teaching in a standards-based world: Finding a sensible balance. Science & Education, 11(4), 405–417.

DeSouza, S., Boone, W., & Yilmaz, O. (2004). A study of science teaching self-efficacy and outcome expectancy beliefs of teachers in southern India. Science Education, 88(6), 837–854.

Education Week. (2000). Unmet promise: Raising minority achievement. Education Week, 19(27), 1,18–19.

Education Week. (2001). Gaining ground. In Pew Charitable Trusts/ Quality Counts. A better balance: Standards, tests and the tools to succeed (pp. 33–40). Washington, DC: Author.

Firestone, W., & Mayrowetz, D. (2000). Rethinking "high stakes": Lesson learning from the United States, and England and Wales. Teachers College Record, 104(4), 724–749.

Gregory, K., & Clarke, M. (2003). High stakes assessment in England and Singapore. Theory into Practice, 42(1), 67–74.

Goodnough, A. (2001, June 9). School board in new dispute on test scores. New York Times.

Kahle, J. B. (1997). Systemic reform: Challenges and changes. Science Educator, 6, 1–6.

Kahle, J. B. (1999, February). Discovering about *Discovery*: The evaluation of Ohio's systemic initiative. Invited address, Fourth Annual NISE Forum: Evaluation of systemic reform in mathematics and science. Washington, DC: National Institute for Science Education.

Kahle, J. B. (in press). Systemic reform: Research, vision, and politics. In S. Abell & N. Lederman (Eds.), Handbook of research on science education. Mahwah, NJ: Lawrence Erlbaum Associates.

Kahle, J. B., Meece, J., & Scantlebury, K. (2000). Urban African American middle school science students: Does standards-based teaching make a difference? Journal of Research in Science Teaching, 37, 1019–1041.

Keeves, J. P., & Alagumalai, S. (1998). Advances in measurement in science education. In B. J. Fraser & K. G. Tobin, International handbook of science education (Part 2) (pp. 1229–1244). Dordrecht, The Netherlands: Kluwer Academic.

Lee, J. (2002). Racial and ethnic achievement gap trends: Reversing the progress toward equity? Educational Researcher, 31(1), 3–12.

Linacre, J. M. (2005). Winsteps Rasch analysis software. PO Box 811322, Chicago IL 60681-1322, USA.

National Commission on Excellence in Education. (1983). A nation at risk: The imperative for educational reform: A report to the Nation and the Secretary of Education. Washington, DC: U.S. Department of Education.

National Council of Teachers of Mathematics (NCTM). (1989). Curriculum and evaluation standards for school mathematics. Reston, VA: Author.

National Research Council (NRC). (1996). National science education standards. Washington, DC: National Academy Press.

National Science Foundation. (1998). The National Science Foundation's Statewide Systemic Initiatives (SSI) Program: Models of reform of K-12 science and mathematics education. Washington, DC: Author.

O'Day, J. A., & Smith, M. S. (1993). Systemic reform and educational opportunity. In S. H. Furhman (Ed.), Designing coherent education policy (pp. 250–312). San Francisco, CA: Jossey-Bass.

Olson, L. (2001). Overboard on testing? In Pew Charitable Trusts/Education Week Quality Counts, A better balance: Standards, tests and the tools to succeed (pp. 23–30). Washington, DC: Education Week.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedagogiske Institut. (Reprinted by the University of Chicago Press, Chicago, 1980).

Rehfeldt, T. (1990). Measurement and analysis of coatings properties. Journal of Coatings Technology, 62, 53–58.

Rentz, R. R., & Bashaw, W. L. (1977). The national reference scale for reading an application of the Rasch model. Journal of Educational Measurement, 14, 161–179.

Sadler, P. M. (1999). The relevance of multiple-choice testing in assessing science understanding. In J. J. Mintzes, J. H. Wandersee, & J. D. Novak (Eds.), Assessing science understanding: A human constructivist view (pp. 251–278). San Diego, CA: Academic.

Scantlebury, K., Boone, W., Fraser, B. J., & Kahle, J. B. (2001). Design of an evaluation tool to measure long-term, systemic reform in science education. Journal of Research in Science Teaching, 38, 646–662.

Thurstone, L. L. (1928). Attitudes can be measured. American Journal of Sociology, 33, 529–554.

U.S. Department of Education, Office of the Deputy Secretary. (2004). No child left behind: A toolkit for teachers. Washington DC: Author.

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis. Chicago: MESA Press/University of Chicago.

Wright, B. D., & Stone, M. H. (1979). Best test design. Chicago: MESA Press/University of Chicago.