

A CASE STUDY OF VERTICALLY MODERATED STANDARD SETTING FOR A
STATE SCIENCE ASSESSMENT PROGRAM¹

Chad W. Buckendahl

Buros Center for Testing/University of Nebraska, Lincoln

Huynh Huynh

University of South Carolina

Theresa Siskind

Joseph Saunders

South Carolina Department of Education

¹ A version of this paper was originally presented at the invited symposium “*Vertically Moderated Standards: Assumptions, Case Studies, and Applications to School Accountability and NCLB Adequate Yearly Progress*” at the annual meeting of the National Council on Measurement in Education, April 2004, San Diego, CA. The symposium was co-listed with the American Educational Research Association and was also sponsored by the AERA Large-scale Assessment special interest group.

Abstract

Under the Adequate Yearly Progress (AYP) requirements of the No Child Left Behind (NCLB) legislation, states are currently faced with the challenge of demonstrating continuous improvement in student performance in reading and mathematics. Beginning in 2007-08, science will be required as a component of NCLB. This paper describes South Carolina's elementary science assessments and its approach to setting achievement levels on those tests. Descriptions are also provided for how the state developed a system of vertically-moderated standards across the range of grades covered by the tests. Included in the process are standard setting activities, Technical Advisory Committee (TAC) deliberations, State Department of Education input, and the recommendation to the state's Board of Education. Recommendations for practice are also provided.

From Content Standards, Through Technical Advisory Committee, to State Board of Education: A Case Study Based on South Carolina's PACT 2003 Science Assessments

Introduction

Under the Adequate Yearly Progress (AYP) requirements of the No Child Left Behind legislation (NCLB, 2001), states are faced with the challenge of demonstrating continuous improvement in student performance in reading and mathematics. Attaining goals that are defined by benchmark and subsequent levels of performance and then factored into a school accountability system are critical components of the federal legislation. Beginning in 2007-08, science will be the third content area with required assessments under NCLB. The challenge of interpreting changes in student performance from one grade to the next has been discussed with a greater sense of urgency since the passage of NCLB.

The process of expressing test scores from several successive grades onto a common scale is not new. Almost from the dawn of testing, it has been carried out under the grade equivalent umbrella for many subjects taught in elementary schools. Using the context of the Iowa Tests of Basic Skills (ITBS; Hieronymus & Hoover, 1986), for example, Petersen, Kolen, and Hoover (1989) indicate "the content represented in all levels of a test from an elementary achievement test battery can be viewed as defining a *developmental continuum* for a particular area of achievement." (p. 231). A developmental scale may therefore be appropriate for this situation. The construction of such a scale requires a statistical process called vertical equating (Slinde & Linn, 1977). More contemporary writers prefer the term "vertical scaling" or "vertical linking."

Test developers are quite successful in constructing developmental scales in the subject areas of reading and mathematics for grade spans such grades 3-8 thanks to the substantial content overlap across grade levels. Unlike reading and mathematics, however, science curriculum and instruction may become more grade level specific in junior high and high school. Science content at these levels may transition from earth science to physical science, to biology, to chemistry, to physics and beyond with little overlap in content or a common underlying dimension that would support linking performance across grades or courses.

Mislevy (1992) and Linn (1993) discuss types of linking methods that seek to add a level of comparability across different assessments. Each author discusses the different levels of comparability in terms of the strength of the link that a given method provides. Practically, the question remains about how to interpret these data in the context of the political questions that need to be answered. As an extension of this linking research, Lissitz and Huynh (2003) report on a strategy that they describe as "vertical moderation" of standards. One example of this is the standard setting process used for the South Carolina PACT 1999 assessments in English language arts (ELA) and mathematics for grade 3 through 8. In this approach, standards were set for endpoint

grades using a common set of performance level definitions for the achievement levels and then interpolating achievement level values for the intermediary grade levels (Huynh, Meyer, & Barton, 2000). Considering these existing options served as the impetus for this study.

This paper uses a case study approach to describe the method used by the South Carolina Department of Education (SCDE) to set achievement levels for the elementary science assessments (grades 3-6). We chose to focus on grades 3-6 for this study because the science content in the elementary grade levels may be more coherent than the more course-specific science content that tends to appear beginning in junior high. Note: both South Carolina and NCLB view grade 6 as middle school. NCLB has three grade designations – 3 through 5, 6 through 9, and 10 through 12. Descriptions are provided for the assessments, the standard setting activities, Technical Advisory Committee (TAC) deliberations, and the presentation to the state's Board of Education. Recommendations for practice are also provided.

Information about PACT Science Assessments

South Carolina's Palmetto Achievement Challenge Tests (PACT) in English language arts, mathematics, science, and social studies are part of the state's assessment system and are used to measure students' performance on the state content standards in these areas. All South Carolina public school students in grades 3-8 participate in PACT assessments. Both individual and school level provisions are included in the state's accountability system that uses PACT scores. These provisions are detailed in the South Carolina Education Accountability Act of 1998 (SC State House, 1998). In summary, students who perform below standards are identified for additional assistance and participate in an academic plan designed to bring them up to the level expected by the standards. Schools receive two ratings under the accountability system, one for the percent of its students meeting standards and one for the level of improvement of its students. South Carolina includes both science and social studies in their state accountability system. Thus, the challenge of articulating students' performance across grade level is current with respect to the state's accountability system and proactive in terms of future NCLB requirements.

The PACT science assessments are intended to provide information on the extent that students in grades 3 through 8 have attained knowledge and skills in inquiry, life science, earth science, and physical science. The assessments are based on the curriculum frameworks and content standards approved by the State Board of Education. One objective of administering the PACT science examinations is to classify students into one of four achievement levels: Below Basic, Basic, Proficient, and Advanced. The operational test forms were developed by the SCDE in conjunction with its contractor Data Recognition Corporation (DRC), which provided test construction, administration, scoring, and reporting services.

The PACT science assessments include a combination of multiple-choice and constructed response items that assess the specified strands of the science curriculum. All multiple-choice items are scored dichotomously. Each constructed response item is scored on a two, three, or four point scale based on a rubric developed by SCDE.

Overview of Standard Setting Method

The recommended range of cut scores that distinguish achievement levels was based on a variation of the bookmark method (Lewis, Mitzel, & Green, 1996; Mitzel, Lewis, Patz, & Green, 2001). This method uses expert judges to examine items on the test and estimate how a typical student on the border between two levels of proficiency will likely perform on that item. Items are ordered empirically from least difficult to most difficult and compiled into a booklet. For PACT 2003 standard settings for science, item difficulties, used to order the items, were estimated from the field and operational test administrations using a one parameter (Rasch) model for multiple-choice items and a one parameter (Rasch) partial credit model for constructed response items. In this application, multiple-choice items were located at the point where students had a .67 probability of success on the item (Huynh, 1994, 1998). Constructed response item score points were located where the upper cumulative probability was .67 of achieving that score point or higher. Over 50,000 student responses were used to establish item difficulty values.

The standard setting process began with panelists working first in small groups of 4-5 persons and then consolidating their discussions to establish a common description of target performers (i.e. those who were “barely” in the respective level) at each achievement level based on performance descriptions. Using the performance level descriptors (PLD) as a reference point, each panelist started with the easiest item and moved through the ordered item booklet until the panelist finds the place where the “barely” Basic student would likely (with at least a .67 probability) answer items up to that point correct. When the panelist places a bookmark at that point, he or she is distinguishing between Below Basic and Basic performers. This process was repeated for each achievement level to distinguish the Basic from the Proficient performers and the Proficient from the Advanced performers. Panelists were then provided feedback in terms of the panel’s median bookmark placement, the range of placements, and the impact of the median placement based on cumulative performance data. Before panelists were shown the impact data, facilitators led a discussion of the panelists’ rationale for their initial bookmark placements relative to the performance level descriptors and the discussion of the target students at each achievement level. This discussion focused on the content represented by the item and how it reflected the likely performance of the target student. After seeing the impact data and having a limited discussion of how to interpret it, panelists made their second round bookmark placements. Panelists were told that the second round bookmark placement could be the same as or different from their first round placement. The cut score recommendations are based on the second round bookmark placements. A given cut score is determined for each panelist by translating the ordered item booklet page number into the corresponding theta location and finding the median values across the panelists.

Performance Level Descriptors for PACT Science

The following performance level descriptors (policy definitions) broadly define the achievement levels into which students in South Carolina are classified on the PACT

science assessments. They are similar to those used for the 1999 PACT assessments in English language arts and mathematics (Huynh, Meyer, & Barton, 2000; p. 38).

Below Basic	This student has not met expectations for student performance based on the curriculum standards approved by the State Board of Education.
Basic	This student has met minimum expectations for student performance based on the curriculum standards approved by the State Board of Education.
Proficient	This student has met expectations for student performance based on the curriculum standards approved by the State Board of Education—this performance level represents the long-term goal for student performance in South Carolina.
Advanced	This student has exceeded expectations based on the curriculum standards approved by the State Board of Education.

Standard Setting Workshop

The standard setting workshops were conducted in Columbia, SC, July 28-29, 2003 for the purpose of soliciting recommendations from a statewide group of panelists outside the SCDE. Evaluations of the workshops were also conducted to measure the panelists' experiences with the process.

Panels and General Introduction

Panels were grouped into grade level teams, 3rd/4th grade, and 5th/6th grade with an equivalent number of educators from each grade level on the team. Each panel provided judgments on both grade level assessments. The rationale for having educators from both grades represented on the team was to facilitate the articulation of standards across the grade levels. The workshops for setting cut scores used panels of 16 and 15 for the 3rd/4th grade and 5th/6th grade teams, respectively (Jaeger, 1991; Raymond & Reid, 2001). The 31 educators across the two panels had an average experience of 17.7 (median = 20.0) years at grade level and content area. These panels were selected and recruited by the SCDE to ensure representation based on geographic location, gender, ethnicity, and district's socioeconomic characteristics.

The standard setting workshop began for both panels with an introduction that included the purpose of the workshops, the role of the facilitators, an overview of the bookmark method and the standard setting process, and the goals of the workshops. After this initial orientation to the process, the two panels broke into their grade level teams.

Detailed Elaboration of Performance Level Descriptors

For each panel, facilitators began by presenting the performance level descriptor(s) (PLD) to the panelists and reviewing the standard setting process. Each of the panels was then further subdivided into smaller groups of four to five panelists within their breakout rooms for an initial discussion of the knowledge, skills, and abilities of target students. Panelists discussed the distinctions that differentiated students' performance for each of the four achievement levels. Using input from these smaller groups, the entire panel established a consolidated understanding of the distinctions between below basic and basic, basic and proficient, and proficient and advanced performances. These discussions were transcribed, printed, and copies distributed to panelists so that they could refer to them during their operational judgments. This discussion separately for each respective grade level's standard setting judgments.

Practice Activities

After discussing the PLDs and distinctions among levels, panelists engaged in a practice activity that allowed them to use a practice set of items to place the bookmarks that separated the four achievement levels. Panelists repeated the bookmark placement process for each of the three cut-points—Below Basic/Basic, Basic/Proficient, and Proficient/Advanced. By placing their bookmark in a given location, panelists were judging that the items before each bookmark represented content that students at that achievement level have a high probability (at least .67) of answering correctly. For constructed response items, the bookmark indicated that the target student was likely to obtain the score point indicated on items placed before the bookmark. The panelists discussed the median and range of bookmark pages established by the panel, anchoring their discussion to the PLDs and achievement level distinctions elicited in the group's earlier discussions. The panelists were also shown impact data reflecting the approximate percent of previous test takers who would have been classified at each achievement level when using the group's initial median bookmark placement.

First Rounds of Bookmarks

Panelists then took their respective grade level assessment without the answer key. This exposed panelists to the range of content and item difficulty found in the item bank. Following this activity, panelists engaged in the first round of bookmark placements using the operational ordered item booklets. After round one, feedback data were provided to the panelists in terms of median page placement for each panel's respective cut point, range of page placements, discussion of bookmark placements, and impact of the median page placement. The discussion of the bookmark placements focused on the content represented by the items around the panel's initial recommended bookmark placement (median), at the higher end of the range of initial recommendations, and at the lower end of the range of initial recommendations. Panelists were asked to characterize their bookmark placements with respect to the content represented by the bookmark relative to the performance level descriptors that had been discussed and transcribed for the panelists earlier. These discussions often reflected educators' experiences with their students who had skills represented by the various achievement levels defined by the performance level descriptors.

3	-1.99	.52	99.6	.76	.41	23.2	1.99	.48	2.0
4	-.55	.35	81.4	.49	.34	30.8	1.58	.40	4.6
5	-.49	.41	74.6	.57	.35	21.4	1.28	.37	4.1
6	-.30	.57	76.4	.29	.31	43.0	1.19	.34	6.6

The panelists' final recommended achievement levels, though, were based on their Round 2 bookmark placements. These placements were made after receiving the performance data described above. The Round 2 recommended cut scores and other assorted statistics are shown in Table 2.

TABLE 2. Round 2 Median Cut Scores, Bookmark SEs, and Percent At or Above Each Cut Score

Grade	Basic Cut Score			Proficient Cut Score			Advanced Cut Score		
	Median	SE	% At or Above	Median	SE	% At or Above	Median	SE	% At or Above
3	-.97	.42	89.9	.85	.37	18.9	1.62	.42	4.6
4	-.70	.35	85.3	.75	.34	21.9	1.61	.41	4.6
5	-.49	.39	74.6	.61	.34	21.4	1.22	.36	4.1
6	-.46	.31	80.2	.29	.30	43.0	1.12	.33	6.6

The impact of the Basic cut score ranged from approximately 75%-90% of students being classified at or above basic across grade levels. The impact of the Proficient cut score ranged from approximately 19%-43% across grade levels. Grade 6's results were at the high end of the range for the Proficient cut score. Because the panels that recommended these cut scores were grouped in a 3rd/4th grade team and a 5th/6th grade team, it might be hypothesized that the impact across the grades in the grade level team would be similar. Although this phenomenon was evident for the 3rd and 4th grade, it was not observed in the 5th/6th grade panel. The impact of the Advanced cut score ranged from approximately 4%-7% across grade levels. Again, there was greater consistency observed in the 3rd/4th grade panel than was observed in the 5th/6th grade panel.

The Round 2 results were taken as the cut scores recommended by the panelists of the standard setting conference. Along with other statistics, they were presented to both the Technical Advisory Committee (TAC) and South Carolina Department of Education (SCDE.) These ranges were also displayed graphically at the median and ± 1 , ± 2 , and ± 3 bookmark SEs from the median for each cut score.

TAC Presentation and SCDE Deliberations

On July 8 and 9, 2003 (about one week after the standard setting workshops), data were presented to members of the TAC and SCDE in Columbia, South Carolina. At this

meeting the standard setting contractor described the procedures from the workshops and presented the results of the workshops. Specifically these results were the median panel-recommended achievement levels, a range of values that were within one, two, and three bookmark standard errors, impact data in terms of the percentage of students at or above each cut score, and the results of the workshop evaluations.

The TAC agreed that the final cut scores should result in impact data that (a) are relatively stable across grade levels, (b) display an across-grade trend line that is similar to other PACT assessments, (c) are consistent with PACT results in mathematics, and (d) supported by national and state data on related subject areas such as NAEP science and the previous state assessment in science (BSAP). Finally, adjustments, if any, to the panelists' recommended cut scores are to be made within the range of three standard errors of the judgments from the bookmark standard setting process.

Pertinent to SCDE deliberations are the data of Table 3 and 4. Table 3 provides the range of impact data (i.e. percent of students at or above the cut score) defined at the end point of the panelist recommended cut (PRC) scores plus or minus one to three SEs. Table 4 provides a summary of the external NAEP, BSAP, and PACT data used in the final SCDE decision-making process. Using these data, the SCDE technical staff worked out a tentative set of final cut scores and relayed them to the TAC for comments and suggestions. These were then taken as official cut scores for the PACT science assessments and presented to the SC State Board of Education for information. Table 5 presents the impact data of the official cut scores.

The report to the Board (SCDE, 2003) includes the following points regarding the SCDE decision on the official cut score for the PACT 2003 science assessments.

- (1) Panelists were informed that their role was advisory and that their recommendations would be reviewed by the Technical Advisory Committee (TAC) and a final decision would be made by the Department.
- (2) The TAC reviewed the panel's recommendations and was struck by the inconsistencies across grade levels. The following considerations were suggested and employed in determining the cut scores.
 - Cut scores should result in relatively consistent results across grade levels within a subject.
 - Science scores are and should be more closely related to mathematics scores.
 - Score patterns should be supported by collateral data including PACT math and ELA from 1999 and 2003, NAEP science and history scores in grades 4 and 8, and BSAP science. The proposed cut scores tend to follow the patterns of NAEP at a less demanding level. The scores are generally more demanding than BSAP.
- (3) Although the committees were not always consistent, cut scores should fall within a two standard-error band around the committee recommendation whenever feasible.

TABLE 3. Range of Percentage of Students At or Above Each Cut Score

Grade	Interval	Percent At or Above Basic Cut Score Range	Percent At or Above Proficient Cut Score Range	Percent At or Above Advanced Cut Score Range
3	PRC +/- 1 SE	75 to 96	9 to 33	2 to 9
	PRC +/- 2 SEs	56 to 99.1	5 to 44	0.6 to 19
	PRC +/- 3 SEs	38 to 99.6	2 to 62	0.3 to 33
4	PRC +/- 1 SE	72 to 94	11 to 36	1 to 9
	PRC +/- 2 SEs	57 to 97	7 to 52	0.4 to 22
	PRC +/- 3 SEs	41 to 99	3 to 68	0.2 to 36
5	PRC +/- 1 SE	58 to 87	11 to 36	2 to 11
	PRC +/- 2 SEs	36 to 96	4 to 52	0.3 to 26
	PRC +/- 3 SEs	17 to 99	2 to 69	0.2 to 41
6	PRC +/- 1 SE	68 to 92	24 to 58	3 to 13
	PRC +/- 2 SEs	48 to 97	13 to 76	0.8 to 28
	PRC +/- 3 SEs	33 to 99	7 to 87	0.3 to 43

Note: PRC = Panelist recommended cut score

TABLE 4. Summary of External Data

Assessment	Percent At or Above Basic	Percent At or Above Proficient	Percent At or Above Advanced
NAEP Science 1996 Grade 4	67	29	3
NAEP Science 1996 Grade 8	61	29	4
NAEP Science 2000 Grade 4	66	29	4
NAEP Science 2000 Grade 8	61	32	4
PACT Math 1999 Grade 3	56	18	5.3
PACT Math 1999 Grade 4	55	17	4.6
PACT Math 1999 Grade 5	53	16	4.4
PACT Math 1999 Grade 6	53	16	4.5
BSAP Science 1998 Grade 3	64*		
BSAP Science 1998 Grade 6	52*		
BSAP Science 1998 Grade 8	44*		

(*) Note: These are the percents of passing students.

TABLE 5. Impact Data of Official Cut scores for PACT Science Assessments

Grade	Percent At or Above Basic	Percent At or Above Proficient	Percent At or Above Advanced
3	56.2	23.2	6.5
4	57.1	21.9	6.5
5	58.2	21.4	7.9
6	58.4	20.1	6.6

Discussion

Although informed by judgmental strategies, the ultimate outcome of standard setting activities remains a policy decision that needs to support the purpose of the assessment system. If assessment systems are intended to be complementary to each other then the results of these systems should convey the uniqueness of each system's purpose. Systems that cannot be distinguished from one another may reflect redundancy in the assessment systems that can be revised for greater efficiency. It appears simplistic to suggest that the results for the program at the heart of this study should fall somewhere between a more stringently defined program and less stringently defined program so the results are meaningful. Many may also question why a standard setting study was even needed in these situations if the parameters of the results were already defined a priori. Although the policy question may be addressed, the psychometric questions linger.

From a measurement perspective, when judgments from standard setting methods fall within an expected range relative to other assessments of similar content, it provides confirmatory validity evidence for the program. We know that different tests have different purposes and different definitions of performance. Thus, the results from these should align with each other representing each test's contribution to the assessment system as a whole. Second, conducting standard setting studies, alignment, and opportunity to learn studies provide information about whether there is a coherent curriculum and consistent instruction across grade levels. In this study we observed some variation in the results across grade levels that might be explained in part by the differences in curricular content and emphasis in science. Third, these activities allow programs to examine whether or not the construct is or can be scaled across grade levels. In this application, there was an attempt to vertically moderate scales in a content area for which it may or may not be conducive. Because of greater course specificity beginning in grade 7, we chose to focus on the grades 3-6 science assessments expecting there to be greater coherence across those grade levels.

An additional measurement question that may need to be explored in a future study relates to the challenge of setting multiple cut scores for an assessment. This question revolves around the extent of the inferences that can be drawn from an assessment and could be characterized as sufficiency of information. If there are limited measurement opportunities at each desired level of inference, there will likely be greater fluctuation in the cut scores that are recommended by the panel making moderation across grade levels more tenuous. Because other states and assessment systems are also faced with this challenge, we offer three recommendations for practice:

1. Utilize relevant, additional data sources to guide the final policy decision making process and add greater consistency to the complementary tests that may be in place in the broader assessment system.
2. Operationally define performance across grade levels to support the desired articulation in the recommended performance standards. Anchoring performance level decisions in the content provides greater support for the policy decisions and demonstrates to stakeholders the types of knowledge, skills, and abilities that may be observed at a given performance level.
3. Evaluate inconsistencies in recommended performance standards for evidence that may not support the policy decisions. For example, if one grade level demonstrates a large discrepancy in the recommended performance standard from what was expected given corollary information, there may be curricular, instructional, and/or assessment explanations for the difference that need to be addressed.

Conclusions

The challenge of setting standards that consider the characteristics of the target student and consider the difficulty of the assessment, yet meaningfully articulate across grade levels is nontrivial. In trying to demonstrate growth in a content area for federal accountability purposes, states are faced with balancing the recommendations for performance standards with the policies that flow from them. As described in this paper,

South Carolina did not rely solely on the information gathered from the standard setting workshops for the PACT Science Assessments, but included additional assessment information to guide the ultimate policy recommendation. By considering assessment systems that are designed for different purposes in the decision-making process, it allows for greater consistency among the systems.

References

- Hieronimus, A. N., & Hoover, H. D. (1986). *Iowa Tests of Basic Skills* manual for school administrators. Chicago: Riverside.
- Huynh, H. (1994, October). *Some technical aspects of standard setting*. In Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments of the National Assessment Governing Board (NAGB) and the National Center for Education Statistics (NCES). (pp. 75-93). Washington, DC: Authors.
- Huynh, H. (1998). On score locations of binary and partial credit items and their applications to item mapping and criterion-referenced interpretation. *Journal of Educational and Behavioral Statistics*, 23(1), 35-56.
- Huynh, H. (2003, August). *Technical Memorandum for Computing Standard Error in Bookmark Standard Setting*. (The South Carolina PACT 2003 Standard Setting Support Project). Columbia, SC: University of South Carolina.
- Huynh, H., Meyer, P., & Barton, K. (2000, October). *Technical Documentation for the South Carolina PACT-1999 Tests*. Columbia, SC: South Carolina Department of Education. Retrieved March 27, 2004 from <http://www.myschools.com>.
- Jaeger, R. M. (1991). Selection of judges for standard setting. *Educational Measurement: Issues and Practices*, 10(2), 3-6, 10, 14.
- Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June) *Standard setting: A bookmark approach*. In D. R. Green (Chair), IRT-based standard-setting procedures utilizing behavioral anchoring. Presentation at the National Conference on Large-Scale Assessment, Phoenix, AZ.
- Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, 6(1), 83-102.
- Linn, R. L., & Kiplinger, V. L. (1995). Linking statewide tests to the National Assessment of Educational Progress: Stability of results. *Applied Measurement in Education*, 8(2), 135-155.
- Lissitz, R. W., & Huynh, H. (2003). Vertical equating for state assessments: Issues and solutions in determination of adequate yearly progress and school accountability. *Practical Assessment, Research & Evaluation*, 8(10). Retrieved May 8, 2003 from <http://ericae.net/pare/getvn.asp?v=8&n=10>.
- Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: Educational Testing Service.
- Mitzel, H. C., Lewis, D. M., Patz, R.J., & Green, D. R. (2001). The bookmark method: Psychological perspectives. In Cizek, G. J. (Ed.) *Setting Performance*

- Standard: Concepts, Methods, and Perspectives* (pp. 249-281). Mahwah, NJ: Lawrence Erlbaum Associates.
- National Center for Education Statistics (2001, November). *National science achievement-level results, grades 4 and 8: 1996 and 2000*. Retrieved on March 1, 2004 from <http://nces.ed.gov/nationreportcard/science/results>.
- National Research Council. (1998). *Uncommon measures: Equivalence and linkage among educational tests*. Washington, DC: National Academy Press.
- No Child Left Behind Act of 2001*, Pub. L. No. 107-110, 115 Stat.1425 (2002).
- Peterson, N. S., Kolen, M. J., & Hoover, H. D. (1988). Scaling, norming, and equating. In Linn, R. L. (Ed.) *Educational Measurement* (3rd Ed.) (pp. 221-262). New York: American Council on Education and Macmillan Publishing Co.
- Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. J. Cizek (Ed.) *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 119-157). Mahwah, NJ: Lawrence Erlbaum & Associates, Inc.
- Slinde, J. A., & Linn, R. L. (1977). Vertically equate tests: Fact or phantom? *Journal of Educational Measurement*, 14(1), 23-32.
- South Carolina Department of Education (1998). *BSAP: 1998 Results of the Basic Skills Assessment Program*. Retrieved on March 1, 2004 from <http://www.myschools.com/reports/BSAP/bsapdata.htm>
- South Carolina Department of Education (2003). *Report presented to the State Board of Education, August 27*. Columbia, SC: Author.
- South Carolina State House (1998). *Education Accountability Act of 1998*. Retrieved on March 1, 2004 from <http://www.scstatehouse.net/code>

