



## FROM THE PRESIDENT

By David A. Frisbie, University of Iowa

In the June newsletter, I indicated that I would provide quarterly reports to members regarding the organizational changes within NCME that will occur over the 2004-05 year. This is my second installment of reports. In addition, I want to share some news about other on-going NCME activities.

**Organizational Changes.** Here is an update on the progress the Board of Directors has made in ensuring that the benefits of NCME membership will be available in the future with high quality service. (Please see the column in the last Newsletter for background on why these services will no longer be available through the AERA central office.)

- **Publications outsourcing.** In our July meeting, the NCME Board voted to pursue a contract for publishing our two journals with one of the five publishers who previously had provided proposals. At the time of this writing, contract negotiations are nearly completed. By the time you read this, I expect a contract will be finalized and planning will be well underway for beginning this new relationship with the 2005 volume of our journals. In the next Newsletter, I'll provide more detail about our new agreement and what it means for NCME. And, of course, I'll indicate who our new publisher is.
- **Management services contract.** During its summer meeting, the NCME Board also heard presentations from three management companies that had submitted proposals for providing central office services. All three are reputable firms that would serve the interests of NCME well. Now we are addressing follow-up questions and seeking additional information. The goal is to make a selection at the October meeting of the Board so that a contract can be finalized prior to our 2005 annual meeting in Montreal.
- **Annual meeting contract.** In late July, the AERA Central Office offered a draft of a proposal for providing annual meeting services to NCME. The Board will review this proposal during the fall, but final decisions will likely be made after the other two contracts mentioned above have been finalized. The goal is to have a contract in place with AERA that will apply to the coordination, planning, and implementation associated with the 2006 annual meeting in San Francisco.

It has been over a year since it became apparent that NCME would need to make major changes in its publications production and organizational management. Certainly it will take more than another year to implement the details involved in the decisions we will make this year. Consequently, there are many individuals to thank for their efforts in getting us to this point—members of the NCME Board this year and last as well as members of the advisory committee convened last year by Suzanne Lane to offer guidance to the Board. The executive committee has handled the details of our work in identifying and working with vendors and in summarizing data from proposals. Thanks to Suzanne for her diligence this year and last and to Jim Impara for his support this year; they've been true colleagues in this process.

**Other Issues.** Planning is well underway for our *annual meeting* in Montreal next year. I encourage each of you to monitor the NCME website for information on international travel and any last minute changes in regulations that might affect your coming and going to Montreal without a hitch. The slate of invited sessions offers a program that promises to be relevant and appealing to all segments of our members' interests. The highlights will be printed in the fall issue of EM:IP.

The **Training** and Professional Development Committee has identified more than a dozen sessions to offer during the days preceding our annual meeting. These include topics not offered for training in recent years and, consequently, more variety than usual should be available. Check the NCME website this fall for details about sessions and procedures for on-line registration. There are minimum and maximum registration limits for these sessions, so don't wait until you arrive in Montreal to register. Your session of choice may have been cancelled or enrolled to its cap.

The **Membership** Committee has begun its work of addressing the recent decline by trying to reconcile the membership database provided by the central office. They have observed that there have been over a thousand lapsed memberships between January, 2000 and May, 2004. Though some of these could be expected to be accurate changes, the number is suspiciously large. If our records show that your membership has lapsed, you likely will be contacted by a committee member for verification. We need to determine whether we have actually lost members (and if so, why) or whether our recordkeeping procedures are faulty. The committee has outlined a number of promising initiatives for increasing our membership level.

There is good news, also, with respect to NCME *finances*. Our financial statement for the year ending June, 2004 (FY 2004) shows a balance of about \$22,400. This compares with deficits of \$45,300 last year and \$107,200 for the previous year. With the changes outlined earlier with new publication and management contracts, it is not possible to make accurate projections yet for the current year (FY2005), but all indications are that we should finish the year with an even larger balance than we did this past year. This turnaround means we can improve on our declining net worth, but more importantly, we can budget for committee activities and new initiatives that have been stalled due to financial constraints in recent years.

I'm convinced that the changes I've described above in publications and management will bring continued stability to NCME and lay the foundation for significant growth. It's time for us to give more serious thought to that growth—how much we'd like to see and what directions it should take. How could member benefits be improved? What might we offer others who work in the measurement and assessment arena to attract them to NCME? It's time for us to take a serious look inward, maybe for a formal, internal review. Let me know (dfrisbie@uiowa.edu) your thoughts and concerns about any of these issues.

## EDUCATIONAL MEASUREMENT (FOURTH EDITION)

By Robert L. Brennan, University of Iowa, Editor

In the summer of 2002, under the initiative of Linda Crocker, then President of NCME, plans began to be formulated for the fourth edition of *Educational Measurement*. At the request of NCME and the American Council on Education (ACE), Robert Brennan agreed to serve as editor of the volume. The contract that formalized publication conditions for the fourth edition was signed in October, 2002, by representatives of Praeger Publishers (a division of Greenwood Publishing), ACE, and NCME. The contract specified that the length and format of the fourth edition would be approximately the same as the third edition--about 600,000 words typeset in two columns on 8.5" X 11" pages.

During late summer and early fall of 2002, Brennan formed an Editorial Advisory Committee consisting of Michael Baer (ACE representative), Lloyd Bond, Linda Crocker (NCME President), Fritz Drasgow, Michael Kane, Robert Linn (editor of the third edition), William Mehrens, Cynthia Schmeiser, and Wendy Yen. In October, 2002, the Committee met at the ACE headquarters in Washington, DC, for a one-day meeting to advise Brennan on the selection of chapters, senior authors, and potential reviewers. Also present at this meeting, and representing ACE, were Wendy Bresler and Mary Swarna. Susan Slesinger represented Greenwood Press.

By the conclusion of the meeting, it was decided that the fourth edition would include an introductory chapter by the editor followed by 21 chapters distributed among three broad areas: (I) theory and general principles; (II) construction, administration, and scoring; and (III) applications. Chapters would have one of three approximate lengths: 150 double-spaced pages including table, figures, and references (approximately 45,000 words), 100 pages (approximately 30,000 words), or 50 pages (approximately 15,000 words).

By December, 2002, Brennan had secured commitments from all senior authors. Each senior author was given the option of choosing one or more coauthors. By March, 2003, over 60 persons had agreed to serve as reviewers (approximately two per chapter) of chapter outlines and drafts. It is planned that the fourth edition will be published in 2005.

The following is a list of the chapters and authors. The chapter titles are "working" titles that may change slightly. For over half of the chapters, there will be some overlap with chapters in previous editions, but generally the overlap is only as much as required to make such chapters coherent, stand-alone manuscripts. It is planned that previous editions (at least the third edition) will continue to be available in one format or another.

1. *Current Perspectives and Future Directions*  
Robert L. Brennan, University of Iowa

## Part I: Theory and General Principles

2. *Validity*  
Michael T. Kane, National Conference of Bar Examiners
3. *Reliability*  
Edward H. Haertel, Stanford
4. *Item Response Theory*  
Wendy M. Yen and Anne R. Fitzpatrick, Educational Testing Service
5. *Scales and Norms*  
Michael J. Kolen, University of Iowa
6. *Linking and Equating Test Scores*  
Paul W. Holland and Neil J. Dorans, Educational Testing Service
7. *Test Fairness*  
Gregory Camilli, Rutgers
8. *Cognitive Psychology and Educational Assessment*  
Robert J. Mislevy, University of Maryland  
(continued on page 3)

### NEWSLETTER ADVISORY BOARD

ROBERT ANKENMANN, University of Iowa  
JUDY ARTER, Assessment Training Institute  
SCOTT BISHOP, Riverside Publishing  
MICHELINE CHALHOUB-DEVILLE, University of Iowa  
KATIE FISK, Connecticut State Department of Education  
JOAN HERMAN, CRESST/UCLA  
SHARON LEWIS, Council of the Great City Schools  
DUNCAN MACQUARRIE, Harcourt Educational Measurement  
WENDY MCCOLSKEY, SERVE  
HILLARY MICHAELS, CTB/McGraw-Hill  
CAROL S. PARKE, Duquesne University  
S.E. PHILLIPS, Consultant  
BARBARA PLAKE, Buros Center for Testing, University of Nebraska-Lincoln  
DOUGLAS RINDONE, Harcourt Educational Measurement

SUSAN M. BROOKHART, Editor, Duquesne University

Send articles or information for this newsletter to:

Susan M. Brookhart Phone: (406) 442-8257  
2502 Gold Rush Avenue Fax: (406) 442-8257  
Helena, MT 59601 e-mail:  
[susanbrookhart@bresnan.net](mailto:susanbrookhart@bresnan.net)

The *NCME Newsletter* is published quarterly. The *Newsletter* is not copyrighted; readers are invited to copy any articles that have not been previously copyrighted. Credit should be given in accordance with accepted publishing standards.

**Part II: Construction, Administration, and Scoring**

9. *Test Development*  
Cynthia B. Schmeiser and Catherine J. Welch,  
ACT Inc.
  10. *Test Administration, Scoring, Security, and Reporting*  
Allen S. Cohen, University of Georgia  
James A. Wollack, University of Wisconsin
  11. *Performance Assessment*  
Suzanne Lane and Clement A. Stone,  
University of Pittsburgh
  12. *Standard Setting*  
Ronald K. Hambleton, University of  
Massachusetts  
Mary J. Pitoniak, Educational Testing Service
  13. *Technology and Testing*  
Fritz Drasgow, University of Illinois  
Richard M. Luecht, University of North  
Carolina at Greensboro  
Randy E. Bennett, Educational Testing  
Service
- Part III: Applications**
14. *Second Language Testing*  
Micheline Chalhoub-Deville and Craig  
Deville, University of Iowa
  15. *Testing Individuals with Disabilities*  
Lizanne DeStefano, University of Illinois
  16. *Group Achievement Testing in K-12*  
Daniel M. Koretz, Harvard  
Laura S. Hamilton, RAND Corporation
  17. *Testing Individuals in K-12*  
Steven F. Ferrara, American Institutes for  
Research
  18. *Classroom Assessment*  
Lorrie A. Shepard, University of Colorado
  19. *Higher Education and Admissions Testing*  
Rebecca J. Zwick, University of California at  
Santa Barbara
  20. *Monitoring Educational Progress*  
John Mazzeo and Stephen Lazer, Educational  
Testing Service
  21. *Licensure and Certification Testing*  
Brian E. Clauser and Melissa J. Margolis,  
National Board of Medical Examiners  
Susan M. Case, National Conference of Bar  
Examiners.

22. *Legal and Ethical Issues*  
Susan Phillips, Independent Consultant  
Wayne J. Camara, College Board

**MECD SPECIAL ISSUE CALL FOR CONTRIBUTIONS**  
*By Robert C. Chope, San Francisco State University*

Robert Chope (rcchope@sfsu.edu) and Charlie Healy (healy@gseis.ucla.edu) will be the guest editors of a special issue of *Measurement and Evaluation in Counseling and Development* (MECD) that will be devoted to new empirical research on **interest measurement**. MECD is the official journal of the Association for Assessment in Counseling and Education, a division of the American Counseling Association.

Potential topics include but are not limited to:  
interest assessment and low ses persons,  
interests and leisure,  
interests and multiculturalism,  
validation of interest measures, particularly card  
sorts and on-line instruments  
assessment in constructivist career counseling,  
integrating assessments such as the 360 degree  
supervisor into career counseling and  
development programs,  
exploring interest measurement over the life  
span, and  
responsibilities for consequential validity.

Articles are due to editor Patricia B. Elmore (pbelmore@siu.edu) by December 15. Guidelines and directions for submitting articles are in the MECD. You are invited to submit a paper or to recommend someone who is engaged in current research in this area and may have something to offer. We would also welcome additional suggestions for topics.

Manuscripts submitted for consideration should not have been published previously nor be currently under consideration by another journal. Submit 5 copies of the paper and a 50-word abstract, plus a 3 and one half inch diskette, to

Patricia B. Elmore, Associate Dean,  
College of Education and Human Services  
Mailcode 4624, Southern Illinois University  
Carbondale, IL 62901-4624

**FEEDBACK INVITED** – The new electronic newsletter format allows more flexibility of length, layout, and content than did the print version. For example, this issue is longer than usual. Editor Sue Brookhart ([susanbrookhart@bresnan.net](mailto:susanbrookhart@bresnan.net)) welcomes your feedback on the newsletter, especially during this transition time.

**CLASSROOM ASSESSMENT LITERACY OF NBPTS BOARD CERTIFIED TEACHERS**

by Judy Arter, ATI, Rita O’Sullivan, UNC, Rick Stiggins, ATI, Martha Hudson, UNC, and Lindo Iovachinni, WRESA

We reported previously about a study conducted jointly by the University of North Carolina, The Assessment Training Institute, and the Western Regional Educational Service Agency to study the relative classroom assessment mastery of National Board for Professional Teaching Standards (NBPTS) certified and non-certified teachers. This column reports the results.

The study was conducted in four parts: 1) developing, from existing NBPTS standards, a unifying conceptual framework of indicators of accomplished classroom assessment practice; 2) creating instruments from this framework to determine teachers’ level of assessment accomplishment; 3) collecting assessment data on two samples of teachers, one certified and the other not; and 4) analyzing differences between the two groups.

After carefully analyzing 18 National Board Standards documents, as well as accepted standards of sound classroom assessment practice reflected in a sampling from professional literature and the measurement community, researchers found that the NBPTS Standards contained a great deal of consistency in their focus on assessment. These consistencies strongly indicate a vision of classroom assessment that emphasizes the use of assessment to promote student learning (assessment *for* learning) rather than just as a means of recording status of student learning (assessment *of* learning, although both purposes for assessing students are acknowledged as being important.” The *Framework of Indicators of Accomplished Classroom Assessment Practice* that emerged from the first part of the study is summarized in Table 1. A detailed description of these indicators can be obtained in *Review of NBPTS Classroom Assessment Standards*, Assessment Training Institute, December 2002, www.assessmentinst.com.

Instrumentation for the study was developed to measure the indicators in the framework. First, a survey was sent to 750 board and non-board certified teachers in North Carolina. From this, 45 matched pairs of teachers were selected for the sample. Instrumentation for the sample included an assessment log, an interview, and a process for analyzing samples of classroom assessments for quality. Results showed that:

- Independent ratings of classroom assessments provided by interviewed teachers showed that NBPTS certified teachers’ assessments are significantly better than those of the comparison group:
  - The assessments submitted by NBPTS certified teachers were found to yield more accurate results tied to learning goals than assessments submitted by non-board certified teachers;
  - Board certified teachers were *less* likely to include behavior and attendance in their grading systems than non-board certified teachers;
  - The assessments submitted by board certified teachers received significantly higher ratings than non-board certified teachers with regard to avoiding potential sources of bias and distortion.
- NBCTs rated their understanding of classroom assessment higher than non-NBCTs on the survey, but this difference was not supported by the test of assessment knowledge on the survey. On that assessment of knowledge, NBCTs differed only in their reported practice of matching classroom assessments with learning goals.

A full report of the findings can be obtained from Rita O’Sullivan, ritao@unc.edu.

**Table 1. Framework of Indicators of Accomplished Classroom Assessment Practice**

<b>Assessments Reflect Valued Student Learning Goals</b>	<ul style="list-style-type: none"> <li>• Teachers have clear learning goals for students</li> <li>• Learning goals focus on the most important things students need to know and be able to do</li> <li>• Teachers have a comprehensive plan for assessing learning goals over time</li> </ul>
<b>Assessment Processes and Results Serve a Variety of Purposes</b>	<ul style="list-style-type: none"> <li>• Teachers use assessment results to plan instruction</li> <li>• Teachers involve students in assessing their own learning to increase student motivation and achievement. Student involvement includes making learning goals clear to students, student self-assessment and goal setting, and student communication about their learning</li> <li>• Teachers assess to inform others about students</li> <li>• Teachers use student assessment results to understand how to improve their own teaching</li> </ul>
<b>Learning Goals are Translated into Assessments that Yield Accurate Results</b>	<ul style="list-style-type: none"> <li>• Teachers use a variety of assessment methods well</li> <li>• Teachers design assessments with learning goals in mind</li> <li>• Teachers design assessments with purposes in mind</li> <li>• Teachers use assessments that appropriately sample learning</li> <li>• Teachers use assessments that avoid sources of mismeasurement, including bias</li> </ul>

<b>Assessment Information is Managed Well and Interpreted Correctly</b>	<ul style="list-style-type: none"> <li>• Teachers carefully record assessment information and keep it confidential</li> <li>• Teachers appropriately combine and summarize assessment information for reporting (including grades). Such summary accurately reflects learning</li> <li>• Teachers interpret and use standardized test results correctly</li> </ul>
<b>Assessment Results are Communicated Effectively</b>	<ul style="list-style-type: none"> <li>• Teachers effectively communicate assessment results to students</li> <li>• Teachers effectively communicate assessment results to a variety of audiences besides students, including parents, colleagues, and other stakeholders</li> </ul>
<b>Miscellaneous Indicators</b>	<ul style="list-style-type: none"> <li>• Teachers advocate for sound assessments and ethical uses of assessment information</li> <li>• Teachers assist others to assess well</li> <li>• Teachers believe that high-quality, student-involved classroom assessment is an essential part of instruction</li> <li>• Teachers prepare students appropriately for standardized tests</li> <li>• Teachers use technology appropriately to assess students, involve students in assessment, manage and analyze information, and report results</li> </ul>

### **AN ANALYSIS OF ITEM MAPPING AND TEST REPORTING STRATEGIES**

*by Joseph M Ryan, Arizona State University*

A recent NSF-funded study examined features and formats of score reports that make them more or less useful to educators for identifying students' strengths and weakness and for designing, monitoring, and adjusting instructional programs. A review of assessment reporting research literature and an analysis of current assessment reporting practices indicate fairly widespread dissatisfaction with the utility of most score reporting approaches. Field-based educators in South Carolina who are deeply involved in curriculum, instruction, and assessment activities participated in two focus groups to provide insights and suggestions about the content and formats of various score reporting approaches. The first focus group was an inductive session in which participants discussed and responded to two questions. First, what information from statewide assessment reports would be most helpful at the district, school and classroom levels in developing curriculum and in planning instruction? Second, what would assessment reports contain and look like to be most useful at the school and district levels?

The critical information and features of score reports that might make them especially useful were identified from the first focus group and the research review. This information was used to design six prototype score reporting strategies and formats that were reviewed and evaluated in the second focus group. The six score reporting formats that were designed, developed, and then evaluated in this study were as follows.

1. Item Content Objective Mapping with Performance Standards
2. Achievement Performance Level Narrative
3. Strand Achievement Level Reporting for Individual Students
4. Strand Achievement Level Reporting for Groups
5. Reporting Observed, Expected, and Differences in Strand and Item Performance for a Group
6. Reporting Observed, Expected, and Differences in Strand and Item Performance at the Achievement Level Cut Scores

The data analyzed to evaluate the score report formats included observational notes, individual participants' comments recorded on evaluation forms, group generated flip-chart listings of key issues and concerns, and quantitative ratings provided by each participant.

The review of research and practice and the results of the focus group analyses led to consistent advice about maximizing the value of score reports. Some of the numerous suggestions and guidelines generated include the following: score reports should be simple, clear, uncluttered, and concise; print features such as font size, use of bold, etc, are important; jargon and technical language should be avoided; critical information should be highlighted; graphs, charts, and tables should be kept simple and should be explained with text; score information should be related to content standards as explicitly as possible; the finest level of reporting detail that is still reliable should be provided; some form of normative information is useful; and, information about reliability and precision should be provided for all inferences provided or suggested in the report. The value of field testing score reports with their intended audiences through focus groups before finalizing reporting formats is strongly recommended. Conducting future studies that document what teachers and others actually do with score report information seems like the next step in this line of research. The complete report of this study can be found at <http://www.serve.org/publications/downloadables.htm> under the subheading "Analysis of State Test Score Reporting Strategies." The research project was initiated by SERVE in collaboration with the South Carolina Department of Education with funding provided by The National Science Foundation, Award No. REC – 99787977.

## STANDARDIZING THE STANDARDS: IN SEARCH OF UNIFORM GUIDELINES FOR STATE TECHNICAL REPORTS

By Douglas F. Becker, ACT, Inc., and Gregory Camilli, Rutgers University

Annual state assessments, thanks to the No Child Left Behind Act of 2001 (NCLB), have become high-profile, high-impact practices in K–12 education for many states. Individual-student, school, and state score reports are increasingly becoming *de facto* tools for the evaluation of educational progress in the United States; as a consequence, many issues and discussions have arisen regarding the interpretation of psychometric data. It is important for the broader educational measurement community to become more involved in discussions about evaluating the quality of state assessment programs as well as the use of test scores. Measurement experts can help to clarify assessment issues, which in turn may foster better educational and instructional practices. In this article, we propose that one way to accomplish this goal is to comprehensively document the technical characteristics of state assessment programs. Further, such interpretive materials should be relatively uniform across state assessment programs in this country. We believe the time has come for a concerted effort that will result in a uniform set of guidelines that apply to how large-scale state testing programs are documented.

The American public rightly views educational testing as having many purposes. These purposes can be divergent depending on diverse educational goals, values, philosophies, and political inclinations. A similar diversity exists among measurement professionals. For example, while some observers have applauded positive outcomes as evidenced in apparent achievement gains (*e.g.*, Cizek, 2001), others have criticized the effects of testing on instructional practice (*e.g.*, Popham, 2003). However, diversity among stakeholders on the pros and cons of testing should not prevent the adoption of a sensible and practical structure for technical documentation that describes how tests are developed, scored, and interpreted.

Quality control is central to the success of any state testing program. Of the errors that have been reported recently, several have involved inconsistent or incorrectly scored items. It is likely that item flaws or inconsistencies are more frequently reported than other kinds of problems because they are easier to detect once items are released. These kinds of items should be identified and the distortions they create should be removed from score reports. However, problems at the item level are likely to be *less* severe than systematic problems involving scaling, scoring, and equating. In either case, damage is done. But only by assuring accurate measurement can there be any hope of estimating intervention or program effects at the level of the student, school, or state. There is much explaining to do once the effects are measured, but the importance of accurate measurement, especially with regard to annual progress, cannot be overestimated. Yet in regard to state educational reform practices, for example, we do not currently have a very good notion of how broadly (across states) technical measurement issues have affected estimates of annual gains.

In the remainder of this article, we provide a rationale for establishing uniform guidelines for technical documentation of tests and assessments. We then provide an outline of what such documentation might include. This outline is offered as a starting point for a fuller discussion among measurement professionals.

**Rationale.** The technical aspects of state assessment programs are carried out by a limited number of service providers with respect to both the number of highly qualified and trained measurement professionals and the number of companies capable and qualified to deliver a large-scale (*i.e.*, state) assessment program. Resources are increasingly becoming an issue. The demands of a single large-scale assessment program—logistical requirements, security, statistical analyses, reporting, and the like—can be staggering. Rhoades and Madaus (2003, p. 30) pointed out that under the assessment requirements NCLB “an amplified demand for testing services without an appreciable increase in the number of service providers in the short term will intensify time pressures already experienced by the contractors.” Given adequate time and resources, we have no doubt that serious measurement problems could be avoided for the most part. But in the current economic climate, time is short and resources scarce. What can be done, in this case, to avoid what Rhoades and Madaus refer to as the “proliferation of undetected human error in educational testing”?

One plan is to develop a comprehensive set of uniform guidelines for state technical reports. There are two potential benefits to having a set of such guidelines. First, quality control is encouraged by the publication of technical results that have a *consistent* format across states and across years. We say *consistent* and not *identical* because testing programs require flexibility. Second, the publication of important technical details permits stakeholders and measurement specialists to investigate and diagnose potential problems. That is, a state may have access to a much broader array of potential analysts than its testing budget may allow if such standardized information is freely available for broader review and comment. It seems doubtful whether the larger measurement community, beyond the resources provided by a particular state’s contractor, can provide oversight without systematic information on test development and administration. Third, test information that is consistent and comparable across states will enable research that is currently not possible.

Accurate assessment is one of the foundations of NCLB. Without comparative research, it is unlikely that we will be able to synthesize information across states and thereby learn from experience. The central issue here is that there currently exists no consistency across states in how technical documentation is produced and reported. And yet it should be a goal to go even beyond

consistent reporting of technical information. State assessment programs should aspire to provide both consistent technical information and accurate testing results; it is our belief that the former will help facilitate the latter.

This recommendation might seem burdensome and even threatening to some. But consider several practical benefits. The guesswork would be taken out of what to include and exclude from technical reports. State testing divisions could gain additional psychometric capacity, because standardization would provide access to data and information for graduate measurement programs and for professional development. Testing services would utilize templates that are more or less uniform to ensure greater consistency across programs when states employ multiple vendors, not to mention smoother transitions of programs between vendors. Finally, these reports could become key as evidence in legal matters or disputes about the validity of one or more aspects of an assessment program.

**Outline.** Some states scrupulously detail the psychometric characteristics of their tests and assessment program in their technical reports, while other states provide only a broad overview. Moreover, material that is included in technical reports can vary greatly from one year to the next in terms of graphical information, tables of item-level statistics, statistical analyses, and so on. As a starting point for a larger discussion, the following information could be included routinely and systematically – the idea here is that there should be sufficient information provided in the technical report so that an independent entity could reproduce (provided data) the results of the program.

Currently, the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) provide guidelines for documentation. At least 76 of the 92 standards covered in the first four chapters of the *Standards* address some level of documentation. Fifteen standards, additionally, are provided in chapter 6, “Supporting Documentation for Tests.” These standards cover a broad range of topics that include, but are not limited to, availability and interpretability of documentation, rationale for the test, specification of target populations, reliability and validation, and computer-generated score reports. Moreover, the statistical information is described primarily within standard 6.5:

When statistical descriptions and analyses that provide evidence of the reliability of scores and validity of their recommended interpretations are available, the information should be included in the test’s documentation. When relevant for test interpretation, test documents should ordinarily include item level information, cut scores and configural rules, information about raw and derived scores, normative data, the standard errors of measurement, and a description of the procedures to equate multiple forms. (p. 69)

The focus, here, is primarily on statistical procedures and results. While the *Standards* provide the broad principles of what should be reported, there is considerable discretion within these guidelines. The criteria presented below are intended as one example of how the guidelines might be practically implemented. That is, the *Standards* address documentation at a general level – this document presents more of the “nuts and bolts” issues:

- 
1. Population (or sample) descriptions for all analyses
    - a.  $n$ , mean, median, interquartile range
    - b. Standard deviation
    - c. Raw score frequency distribution, cumulative frequency
    - d. Statistical tables (given sufficient  $n$ ) for raw, scale, and theta scores for:
      - i. Achievement level groups
      - ii. NCLB groups (e.g., LEP, Special Education, gender)
      - iii. Achievement level-by-NCLB group
  2. Classical item-level statistics for all items with item format (e.g., MC, OE) indicated
    - a.  $p$ -values, item means
    - b. Item–test correlations
    - c. DIF analyses for NCLB groups, including sample size
    - d. Agreement rates and kappa statistics for OE items
  3. Reliability and validity information
    - a. Hand–scoring reliability and validity
    - b. Rater effects
    - c. Reliability and validity of individual and group scores
    - d. Reliability of classification decisions
  4. Test equating and scaling results
    - a. Description of horizontal equating methods
    - b. Description of vertical equating methods

- c. Delta plots for adjacent years for anchor items, if applicable
  - d. Identification of item format for anchor items, if applicable
  - e. Description (and formulas) of score conversion to scale
  - f. Descriptions of rounding procedures
5. IRT results
- a. IRT model specified for all items
  - b. IRT item parameters for all items
  - c. Item fit statistics
  - d. Test information curves
6. Standard-setting results
- a. Description of standard-setting model and procedures
  - b. Description of judge/expert panels
  - c. Achievement level descriptors and final-round statistics across judges:  
Mean, median, standard deviation, interquartile range
  - d. Achievement level recommendations of judges
  - e. Final achievement level thresholds:  
In raw score units, scale score units, and theta units

Not all assessments will use the same kinds of measurement procedures, and standards should not be seen as an attempt to prescribe methodology. Rather, there are fundamental aspects of measurement that should be included in nearly all technical reports for large-scale testing. The above list of suggested documentation criteria is provided for illustrative purposes. Once a set of specifications is identified, the resulting technical documents should be freely available to the public via the Internet and other appropriate means.

**Conclusion.** NCLB is a major driving force in educational assessment, and this legislation would seem to lead states to have more similar, though broad, goals for their testing programs. The critical need for technical reports across state programs to be based on a comprehensive set of uniform guidelines cannot be emphasized strongly enough. A national panel for this purpose should be convened at the earliest possible moment. The panel should have broad representation including educators, policy makers, and independent, as well as private sector, measurement specialists. It remains to be determined what kind of mechanism could be used for constituting, charging, and funding a national panel to do this work. Certainly the federal government has a compelling interest in the development of technical standards, yet it can be argued that state interests are even greater. Measurement professionals from both the university and private sectors also have an important role to play. The National Council for Measurement in Education and the Council of Chief State School Officers would seem to be a natural partnership for leading such an effort; other relevant agencies might include the American Educational Research Association, the American Psychological Association, and the National Research Council.

Quality control of technical measurement factors is central to the success of any state testing program. The time is right for a comprehensive set of uniform guidelines that apply to how large-scale testing programs are documented. Compared to errors regarding individual test items, scaling and equating inconsistencies are much more likely to confuse state educational policies as well as to muddy the debate on the merits of high-stakes testing. Beyond quality control, accurate test results and thorough documentation will enable states to identify both the benefits and any possible negative consequences of assessment more readily. Uniform guidelines for state technical reports will enable a process by which assessment information can be more accurately used to understand and address state education issues.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice*, 20(4), 19–27.

No Child Left Behind Act, 20 U.S.C. ¶ 6301 *et seq.* (2001).

Popham, W. J. (2003). Seeking redemption for our psychometric sins. *Educational Measurement: Issues and Practice*, 22(1), 45–48.

Rhoades, K., & Madaus, G. (2003, May). Errors in standardized tests: A systemic problem. (Report of the National Board on Educational Testing and Public Policy.) Chestnut Hill, MA: Boston College Center for the Study of Testing, Evaluation and Public Policy. (Available at: <http://www.bc.edu/research/nbetpp/>).