

**REPORT OF THE TECHNICAL ADVISORY COMMITTEE
ON THE
MASSACHUSETTS TESTS FOR EDUCATOR LICENSURE**

**by
William A. Mehrens, Stephen P. Klein, and Robert E. Gabrys**

**Submitted to the
Massachusetts State Department of Education
and
Commissioner of Education**

January 14, 2002

**REPORT OF THE TECHNICAL ADVISORY COMMITTEE
ON THE MASSACHUSETTS TESTS FOR EDUCATOR LICENSURE**

by

William A. Mehrens, Stephen P. Klein, and Robert E. Gabrys

Introduction

The members of the Technical Advisory Committee are pleased to submit this report in response to the request from the Commissioner of Education for a review of the proposed Massachusetts Tests for Educator Licensure. The effort on which Massachusetts is embarking is no small task and clearly affirms its commitment to improved teacher training as a major part of its systemic reform effort. We recognize that there is no assumption that licensure tests will, by themselves, raise the performance of Massachusetts public school students. However, such tests are an essential component of the total reform effort. In our professional opinion, the Massachusetts teacher licensure testing program is a strong, sustainable, psychometrically sound, and essential component of that reform effort.

I. Definition of the Task and Background Information

The Technical Advisory Committee was

"To advise the Commissioner of Education on the optimum balance of test program technical quality and feasibility of future implementation as the Department updates the Massachusetts Tests for Educator Licensure (MA Educator Certification Tests) by reviewing proposed information, materials, and procedures for their technical quality (validity, reliability, and other performance characteristics) and providing a written report" (taken from the RFR).

National Evaluation Systems (NES) is the contractor for the Massachusetts testing program. NES' initial proposed procedures and documents were developed prior to our work and then shared and discussed with us during various meetings. While attention was paid to the past testing practices in Massachusetts, we were not engaged in a critique of those practices and the previous testing program except as those procedures are proposed to continue.

II. Background on Licensure Testing

It is common in the education profession to criticize teacher licensing tests for not doing things they were not intended to do. Therefore, it is important to distinguish the inference that test builders and test users wish to make from the inferences that others may draw, or claim one cannot draw, from the scores.

Professional opinions about the development and validity of a teacher licensure test must be based on an understanding of what such a test is designed to do. The answer to the question of what inference should be drawn from the scores on tests, such as the Massachusetts Tests for Educator Licensure, is essential in determining the validity of the test, because validity has to do with the degree of accuracy of the inference drawn. The inference to be drawn from a teacher competency test is simply that the individual test taker has, or does not have, at least a minimal level of the knowledge that the test is designed to measure. A licensure test is not intended to be a measure of, or a predictor of, performance or effectiveness. It is simply a measure of the knowledge and skills in the domain being assessed.

Like other licensure tests, the Massachusetts Tests for Educator Licensure are not designed as either diagnostic or employment tests. Using licensure tests for either of these two purposes is to misuse these tests. Institutions of higher education involved in training teachers should develop their own diagnostic tools. To provide sufficiently reliable scores for making diagnostic decisions about individual students, these tools would need to contain substantially more questions and take considerably more testing time than what is required for licensure examinations.

The Joint AERA/APA/NCME Standards for Educational and Psychological Testing (1999) (hereafter referred to as the Joint Standards) provides professional guidelines for what constitutes appropriate test development and use. That document states the following with respect to certification (licensure) tests:

Tests used in credentialing are designed to determine whether the essential knowledge and skills of a specified domain have been mastered by the candidate (p. 156)

Validation of credentialing tests depends mainly on content-related evidence, often in the form of judgments that the test adequately represents the content domain of the occupation or specialty being considered. . . . Criterion-related evidence is of limited applicability in licensure settings because criterion measures are generally not available for those who are not granted a license (p. 157).

Thus, the professional standards do not consider it necessary, or even appropriate, to validate scores on licensure tests against an external criterion measure of success on the job.

The question of professional acceptability of any test is not whether scholars can find fault with a process. Any scholar in the field, no matter what test he/she reviewed, could identify ways to improve or criticize the test construction and validation process. An idealized goal is unrealistic because no test would meet it. Consequently, the Joint Standards document states,

Evaluating the acceptability of a test or test application does not rest on the literal satisfaction of every standard in this document, and acceptability cannot be determined by using a checklist. ... Evaluating acceptability involves professional judgment ... (1999, p. 4).

Of course, it is appropriate to identify inconsistencies with these Joint Standards, and we have kept the Joint Standards in mind in reviewing the test contractor's plans for Massachusetts.

We should also make it clear that it is not necessary that things must be done in our preferred fashion to be considered acceptable. At times we will make suggestions regarding how we think something should be done. It does not follow that should NES proceed in a different fashion, that what it does is unacceptable. We recognize that testing is like instruction, a dynamic process subject to constant review and validation. Hence, we recommend a technical advisory committee (TAC) be established to serve in an advisory capacity to the commissioner, and that the TAC be composed mainly of licensure assessment experts.

As noted earlier in this section, while it is crucial to recognize the role of the Joint Standards in developing the tests and in protecting test takers from abuse, it is equally important to understand that the Joint Standards also impose responsibilities on test users. The Commissioner and Department of Education are clear on the intent of the testing program. It is equally important and a heavy burden on them to ensure that test users have adequate information regarding the misuse of licensure test results, such as using them for employment decisions. The Joint Standards are clear about the responsibilities of designers, takers, and users. It might be worthwhile sponsoring some regional meetings on those Joint Standards or having teacher training Institutions of Higher Education (IHEs) do so with their students and participating school districts as part of their implementation plan.

III. Procedures Followed

The expert technical panel participated in four meetings with staff from the Massachusetts Department of Education and National Evaluation Systems, Inc. Prior to each meeting, various materials were shipped to us to study. Subsequent to each meeting we were sent revisions of the plans that NES prepared in response to the discussions and suggestions made at the meetings.

The first meeting was on March 21 and 22, the second on May 23 and 24, and the third on July 24 and 25. A portion of the third meeting was devoted to meeting with four individuals who were selected to represent the Board of Higher Education and the Association of Independent Colleges of Massachusetts. This committee represented the public and independent sectors of higher education in the Commonwealth. We also reviewed a letter and information from Dr. Larry Ludlow of the Lynch School of Education of Boston College regarding the testing program. We held our fourth meeting

via a conference call on November 13. Preliminary to the completion of our final report, we discussed a working draft of it with the Commissioner and Department of Education staff on December 17, 2001.

The materials we initially received were voluminous and included some technical information on previous tests. While we read these, we did not focus our attention (or this report) on the quality of the previous tests. Instead we focused on the plans for the new tests. Hence, it should be noted that the psychometric matters we considered were in the context of the new licensure program, not criticisms of the old. Hence, the main topics of discussion at each meeting were the planning documents that serve as guidelines for the updated and new tests for the "Massachusetts Tests for Educator Licensure." We did some of our own independent analysis of data that were provided by NES to make sure we understood fully the analyses that the contractor had performed.

In our opinion, the original NES planning documents were reasonably good and represented considerable thought. Nevertheless, we found a number of points in these documents that, along with the comments from the higher education representatives, were worthy of further discussion. As mentioned earlier, we had four meetings in which these documents and revisions of them were discussed and alternative procedures were recommended. We found that NES and Department staff were very receptive to our suggestions. We commend both NES and Department staff for their open, professional approach. Indeed, it was their willingness to change procedures in response to our observations and suggestions that resulted in us being convinced that the final documents represent acceptable planning documents. In the Forward of the final document, NES notes that the documents are a "product of extensive review and revision based on meetings of MDE, the Technical Advisory Committee and NES...." We agree that statement is an accurate portrayal of the final document and one that we all support.

IV. Overview of the Issues Addressed

In our meetings with the staff from the Massachusetts Department of Education and NES, we were provided with drafts of the "planning documents" for the Tests for Educator Licensure. The first set of draft planning documents was dated May 2001. Following discussions of the content of these drafts, NES prepared subsequent versions of the working documents dated July 2001. We also had discussions regarding these drafts. We received a third set of drafts of the working documents and a final set of documents dated November 2001.

We will comment briefly on each of the parts of the plan as they are explicated in the November 2001 documents. At times we will comment on the changes as evidence that the joint efforts produced a sound, acceptable set of planning documents. All of our comments are based on the understanding that the Department will be establishing a permanent Technical Advisory Committee to review and advise on testing decisions as the program is implemented.

A. Management Planning Document for New Test Development

This nine-page document provides a well thought out outline of how the new test development will proceed. We believe that a reasonable time line was established and that the process of forming advisory committees was basically sound. Some early discussion focused on the advisability of having members of the Content Advisory Committee invited to the passing score setting activities. We discussed some of the implications of such a process and thought it might be preferable to not have such members on the passing score committee. As we understand it, this original idea has been dropped as a required feature, although there is no explicit statement in the plans that prohibits such membership.

We commend the plans to establish three advisory committees: content advisory committee, bias review committee, and passing score panel. We find the overall plan for selecting the committees to be acceptable. This plan by the Department of Education is another piece of evidence regarding their openness to involve the education community in the development and implementation of the licensure-testing program.

B. Preliminary Test Form Design Document

This document presents a description of the proposed test form designs under three broad headings: Test objectives (the term frameworks was used in the original draft), test form design and blueprint, and test form development and construction.

Again, we felt the Department and NES staffs were responsive to our suggestions. The May draft suggested that "each objective for a field will typically cover about the same amount of content." We felt that such a "requirement" could result in rewriting objectives in a somewhat unusual fashion to make them equally broad or weighting objectives in a way that might not seem sensible. After some discussion this NES proposed procedure was dropped.

We believed that the design regarding the number of multiple-choice items and constructed-response items resulted in a reasonable balance between the two formats. A four-hour test should be long enough to provide for sufficient content coverage and estimated score reliability. The Massachusetts Department of Education, through the work of its committees, will ultimately establish the weights assigned to each section of the test (i.e., multiple-choice and constructed-response). We believe that determination will be done in a deliberate fashion and the process of making that determination will be well documented and involve some of the State test committees outlined above.

Without going into details about the changes between the initial draft and the final one, there is evidence that our preferences were given consideration. We are satisfied that the final draft describes professionally acceptable procedures.

C. Planning Document for Test Objectives Development and Validation

This document presents a discussion of ten tasks involved in defining the content of the Massachusetts Tests for Educator Licensure. These tasks begin with obtaining and analyzing the program and policy materials. They end with the preparation of test specifications. The total document runs 26 pages plus an appendix.

As with the previous documents, this one is thorough and represents a professionally acceptable set of procedures.

D. Planning Document for Test Item Development, Validation, and Pilot Testing

This is a seven-page document describing seven tasks. These tasks include drafting the items, conducting bias and content reviews, developing pilot test forms, recruiting participants, administering and scoring the pilot test, and analyzing the data.

Again, NES was responsive to concerns we raised in previous versions of this document. For example the new test design will include 100 multiple-choice items and 2 constructed-response items. This is an increase from the originally planned 80 multiple-choice items and should result in a test with more reliable scores. The statistical analysis of the pilot items seems adequate. Items will be reviewed if they do not meet certain criteria regarding difficulty, discrimination, and Differential Item Functioning (DIF) statistics.

E. Planning Document for Setting Passing Scores

This document details the following four tasks: establishing the passing score panels, conducting the passing score conference, calculating the panel-based recommended performance levels, and establishing the passing scores.

The panel will use a "modified Angoff" procedure, which is a popular and generally accepted procedure for establishing passing scores in both licensure and large-scale assessments. We spent some time with NES and the Department staff discussing the charge to be given to this panel. The plan is to ask the panel members to answer the following question related to selected response items:

Imagine a hypothetical group of individuals who are just at the level of subject matter knowledge required for entry-level teaching in this field in Massachusetts' public schools. What percent of this group would answer the item correctly?

For the constructed response items, the panel will be asked what score, on a six-point scale, represents the level of response that would have to be achieved for entry-level teaching.

In discussing task 4 (finalizing the passing score), it is stressed that the Commissioner will set the passing score for each test based on input from a variety of sources. That

procedure is sound and follows the practice of many states where the passing score must be determined on the basis of psychometric and other public policy considerations.

Again, we find that the plans proposed in this final November 2001 document represent acceptable practice.

F. Preliminary Test Scoring and Reporting Document

This document provides information related to the scoring of the tests and the reporting structure for results. The Overview of Test Scoring section describes how the multiple-choice item scoring and constructed-response item scoring are accomplished. For the constructed-response items, all are read by two scorers and a third scorer is used if the scores from the first two readers differ by more than one point. Scorer selection criteria, scorer orientation, and scorer training are all well conceptualized and well described in the report.

We concur with the procedures as outlined, with the exception that we believe that when re-scoring of the multiple choice component occurs, it should be done by hand, not a machine re-scoring.

The section on Analysis and Reporting of Results covers Test Form Equating; Scaled Scores; and Reports for Candidates, the Department and Board, and Institutions. An area that presents particular challenges to test developers is comparability of constructed-response items within and across test forms. In the case of the Massachusetts Tests for Educator Licensure, different constructed-response items are typically used on each test form across administrations. Comparability is achieved through a judgmental process that includes marker responses as well as a historical anchor set of marker responses. Scorer orientation, calibration, and monitoring are also described in the document. We believe the NES plans are well considered and generally appropriate. Nevertheless, we would like to see some consideration given to exploring the possibility of scaling the constructed-response items for those fields where the sample size is sufficient for such a scaling process to be appropriate.

The section on scaled scores was the subject of considerable discussion among the TAC, NES personnel, and Department staff. The procedure described on pages 11-15 of the Preliminary Test Scoring and Reporting document is the result of that discussion. We certainly agree that the procedure is professionally acceptable and may indeed be the preferred methodology. However, it may be that for the fields with large numbers of candidates (i.e., 200 or more), a modified z-score or similar method described (but not planned on being used) would be preferable. This is an issue that the contractor should consider and provide a rationale for a final decision regarding the selected procedure to the Department. This rationale should include documentation of the procedures and results of any relevant empirical analyses that NES conducted. We are pleased that NES took this issue seriously and, in fact, ran some data showing that the percent of consistent decisions across the different strategies was always above 90%.

The sub-section on Reports for Candidates, the Department and Board, and Institutions is one area in which we have two disagreements, namely: 1) the reporting of individual scores for those who pass, and 2) reporting sub-area scores by individual. We also recommend that eventual or cumulative pass rates should be reported in addition to initial (first timer) and repeater pass rates.

In our professional opinion, scores should not be reported to those candidates who passed the tests. We believe that just notifying candidates that they passed (i.e., rather than giving them their scores) will help to safeguard against abuse of scores beyond the stated intent of the licensure program.

Individuals are to get their score reports about five weeks after each administration. Details of what will be provided to them are spelled out (see pages 16 and 17 of that section of the report). We do not agree with the plan to include an indication of the candidate's performance on each sub-area in each test. While we understand the desire for information on each sub-area of the test, we do not favor providing such information at the individual level. As we pointed out earlier, licensure tests are not designed to be diagnostic. The reliability of the differences in scores between sub-area scores would simply not be high enough for us to want individuals to have those sub-area scores.

The sub-section on Post-administration summary reports specifies that the percent of candidates passing each test will be reported both for first-time takers and repeaters. We feel that an eventual or cumulative passing rate for all test takers should also be reported.

The document specifies that institutions will receive summary reports for candidates who requested that their results be forwarded to that institution. Again, in that section, it is suggested that institutions will receive sub-area performance by individual candidate. We object to that for the reasons cited above. We would not be opposed to institutions receiving the mean scores of the sub-areas. This would allow for some institutional diagnostics. The mean sub-area scores would be considerably more reliable than the individual sub-area scores.

The document includes a major section on Annual Technical Reporting. This section covers, among other topics, discussions on the level of reporting, what test reliability information will be provided, scorer agreement information, and a pass rate report. This section should also contain the results of the implementation of procedures in low incidence test areas and the results of checks for possible breaches in test security when the same test form is repeated, such as substantial changes in item statistics.

We were advised by Massachusetts that federal requirements require annual reporting of passing rates by institution for program completers. However, such data would be misleading for the basic skills tests if some but not all institutions made admission into their teacher education programs contingent on passing these tests. Hence, it would be useful to also report first-time and repeater passing rates for each administration by institution for all takers, i.e., regardless of whether or not they completed the program. In addition, three-year eventual passing rates should be reported by institution for cohorts of

program completers so that candidates can be better informed about the percentage of graduates from each school that are eventually licensed to teach in Massachusetts.

Summary data should be provided to institutions, i.e., number of takers, means, standard deviations, and frequency distributions for total scores and sub-area scores, provided such reporting does not compromise the confidentiality of individual test takers. These data should not be reported for test areas in which there are fewer than 10 takers. Scores for individual candidates should not be provided by name to the institutions without the consent of those individual candidates.

With the exception of our views on score reporting at the individual and institutional levels, we are pleased with the Preliminary Test Scoring and Reporting Document. It, like the other sections of the report, documents a professionally acceptable plan.

G. Guide for Information Retention

This section contains a detailed explanation of records and information retention as it relates to the total test design, development, implementation, and results process. We find it acceptable as written.

We do, however, recommend the consideration of retention of test administration seating charts by administration at each testing site. One can expect that issues of “cheating” might represent one of the most public aspects of the retention process, and it would be advisable to maintain records on individuals to help adjudicate such situations. Retention for a period of two years is probably sufficient.

We also believe that as much as is feasibly possible, records should be maintained through electronic means, as well as the hard copies of records.

V. Issues and Recommendations

The following constitute the final recommendations of the TAC.

Recommendation 1: We recommend that procedures be established regarding the openness and sharing of data. We anticipate that unless there is a compelling and test compromising reason, test information should be available to the research community without individual or institutional identifiers. Data that are public should be available in an electronic database so that they can be manipulated in ways other than their original presentation.

The Department should adopt a public disclosure stance that places a heavy burden on decisions about non-disclosure. Hence, all technical reports, test administration reports, test development committees’ reports and minutes, and test statistics should be available provided that individual and institutional confidentiality is maintained. It is advisable that an annual technical report be produced, and disseminated upon request. While this

report of each year's test administrations would be technical in nature and not one written for the general public, the documents should be created and available as demonstrable and credible evidence of the integrity of the licensure testing program.

Recommendation 2: We recommend that data on procedures and results presented to the State's teacher training Institutions of Higher Education (IHEs) be in an electronic format. This format should allow each institution to analyze its own data without compromising the confidentiality of the data for identifiable candidates. For example, the school should not be able to ascertain a given candidate's scores unless that student has signed a non-coercive consent form for the release of those data. Provided it would not breach or jeopardize candidate confidentiality or test security, the electronic file should contain each candidate's total multiple choice score, total score on the constructed response section, overall total score, pass/fail decision, and repeater status. IHEs and other organizations that prepare teachers also should receive mean scores for the sub-areas for first time and repeat test takers, provided this does not jeopardize candidate confidentiality.

Recommendation 3: We recommend that a permanent Technical Advisory Committee be established with a charter to advise and to evaluate whether procedures are being implemented as planned. The education community in Massachusetts should view the TAC as a group charged with ensuring that the quality of the testing program always exists. It also advises on procedures and issues that arise relative to discrete test administrations, results, psychometric issues, review of data that cannot be released without compromising the program, and test policy. The TAC should be composed mainly of nationally recognized experts in large-scale licensure assessment who are knowledgeable of the Joint AERA/APA/NCME Standards for Educational and Psychological Testing (1999), and who are not opposed to large-scale entry-level teacher testing.

Recommendation 4: We recommend that consideration be given to scaling the scores on the open-ended sections. While issues of combining both selected and constructed response components of the test always exist, it may be advisable to utilize scaling procedures since all tests will contain at least two constructed response situations. Scaling may be a more appropriate procedure for test forms that are taken by 200 or more candidates than those taken by relatively small numbers of candidates. While the measurement field is not of one mind on this issue, an informed decision based on the merits of each option should be explicated by the contractor to better inform the Department decision, in conjunction with its permanent TAC.

Recommendation 5: We recommend that results be reported for first timers, repeaters, and those eventually passing. In fact, eventual pass rates should be used for much of the reporting of the impact of the testing program. Reporting by test administration rather than in the aggregate will continue to draw attention to those who have failed the test rather than to the expected impact of the program—an assurance of the content knowledge of those seeking licensure in Massachusetts. One can expect variation in test performance from administration to administration. That variability is not the focus of

the program, but rather the assurance that prospective teachers demonstrate content proficiency regardless of the number of attempts. This is especially important in reporting to the State Board of Education on the effectiveness of the program. Hence, the contractor should report summary results by institution and license area after each test administration to the Department for first timers, repeaters, and those eventually passing for its own analysis. Over time, eventual pass rates at the end of a three-year period should be used for judging the impact of the testing program.

Recommendation 6: We recommend that certain scores be reported to the candidates who fail, but not to those who pass the test. Specifically, the scores reported to unsuccessful candidates should include their total score on both the multiple choice and essay portions of each test as well as their overall total score. We do not recommend releasing sub-area scores to individual candidates because these scores probably do not provide a sufficiently reliable basis for determining whether a candidate is actually more proficient in one sub-area than in another (and the tests are not designed to provide that type of diagnostic information). Releasing scores for individuals who passed may be a major contributor to using those scores beyond the intended purpose of identifying individuals who met a content knowledge criterion.

Recommendation 7: We recommend that statistically sophisticated analysis techniques be considered for tests taken by relatively large numbers of candidates. We understand the importance of consistency across test areas, but believe that this consistency should occur on two levels: one for tests taken by small numbers of test takers, and one for tests taken by large numbers that would allow for statistical analyses to be conducted.

Recommendation 8: We recommend that consideration be given to using all 100 items in computing a candidate's multiple-choice score; i.e., the 80 that are now scored plus the 20 that are designated as pilot-test items.

Recommendation 9: We recommend that for tests taken by relatively large numbers of takers (200 or more), early item analysis results should be used to flag items and revise scoring keys as warranted. Item analysis data also should be used in selecting items for subsequent forms, such as by not re-using items with high or low p-values. This is particularly important in low incidence test areas where test forms will not be revised for several years because of low numbers of test takers and the need to develop multiple forms.

Recommendation 10: With respect to accommodations for handicapping conditions, we concur with the procedures outlined by NES. They are reasonable and represent "best practices" in the context that the accommodations are still allowing a measure of the construct to be obtained. We make no reference to alternative testing procedures since those issues lie outside the scope of the licensure testing program and must comply with ADA regulations.

Recommendation 11: With respect to the issue of challenging test scores, we concur with the NES procedure regarding constructed response items. The NES procedure

constitutes a *de facto* review in that scores that are not within the confines of adjacent agreement are automatically re-read by a third reader. With respect to the multiple-choice component, we support the re-scoring procedure that is in place with one adjustment; namely, we recommend the re-scoring be conducted by hand rather than by a computer.

Recommendation 12: We recommend that NES establish a procedure for routinely checking for possible breaches in test security when the same test items are used on more than one occasion. This procedure should include reviewing and flagging items whose statistics change markedly across test administrations. The proposed procedures for doing this should be submitted to the Massachusetts Department of Education for adoption and be considered by the newly appointed Technical Advisory Committee.

Recommendation 13: We recommend that a spirit of openness with the public, institutions of higher education, school systems, and licensure candidates characterize the teacher licensure testing program. We recommend that planning documents used for the program be public with two restrictions as determined by the contractor, Department, and permanent Technical Advisory Committee: 1) if making the material public is determined to compromise the anonymity of individual test takers; and 2) if making certain material public is determined to compromise test security.

**Summary of Qualifications of
Technical Advisors to the Commissioner:
Massachusetts Tests for Educator Licensure**

Robert E. Gabrys, Ph.D.

At present, Dr. Gabrys is Chief of the Education Office of NASA's Goddard Space Flight Center, where he has served in this capacity for the past five years. He has extensive experience in overseeing and administering large-scale, state testing programs for teacher licensure. His prior work has included seven years in the Maryland Department of Education as Assistant Superintendent for Research and Development, Assistant Superintendent for School Performance, and Branch Chief for Program Assessment, Evaluation, and Instructional Support. During his eight years in the West Virginia Department of Education, Dr. Gabrys served as Director of the Office of General and Professional Education, Director of Educational Personnel Development, and Coordinator for Educational Personnel Certification. He also has been a teacher and vice principal in Maryland and Virginia.

Dr. Gabrys earned his doctorate at Syracuse University.

Stephen P. Klein, Ph.D.

Dr. Klein has been a Senior Research Scientist at the RAND Corporation since 1975. In that capacity, he has designed and directed studies in education, criminal justice, health, and military manpower. He has served as a consultant to national and state licensing boards for the past twenty-seven years. As such, he provided advice on testing and scoring practices, racial/ethnic bias, differential item functioning, content and known group validity, reliability of essay grading practices, applicant attitudes, test security, moral character investigation procedures, alternative methods for measuring lawyering competencies, establishing pass/fail standards, and the allocation of testing and reader time. These boards have included the National Conference of Bar Examiners, 25 state bar boards, the American Institute of Certified Public Accountants, the National Board of Professional Teaching Standards, the California Commission on Teaching, the Puerto Rico Board of Medical Examiners, and the Society of Actuaries. He also has served as an expert witness at state and federal legislative hearings and court procedures, including court's technical advisor in a federal class-action suit involving a licensing test for teachers. Dr. Klein was a member of the National Research Council/National Academy of Sciences' Committee on Assessment and Teacher Quality. His prior experience has included seven years at University of California, Los Angeles in evaluation, research design, tests and measurement, and statistical analysis and evaluation of educational programs and criterion referenced tests. Before that, he worked for four years as a research psychologist at Educational Testing Service (ETS).

Dr. Klein earned his doctorate at Purdue University in industrial psychology, with minors in labor law and tests and measurement.

William A. Mehrens, Ph.D.

Dr. Mehrens has been a Professor in the College of Education at Michigan State University for the past thirty-one years. His prior experience has included junior faculty positions at Michigan State and the University of Minnesota. He also served as Assistant Director for Research and Statistics for the National Assessment of Educational Progress (NAEP), a high school counselor, and a junior high school mathematics teacher. His organizational leadership roles in the American Educational Research Association (AERA) include Executive Board Member and Vice President of Division D, Measurement and Research Methodology. For the National Council on Measurement in Education (NCME), Dr. Mehrens served as Member of the Board of Directors, President Elect, President, and Past President. He received the NCME Award for Career Contributions to Educational Measurement in 1997. Dr. Mehrens has written numerous textbooks and other publications on educational and psychological measurement and evaluation, teacher licensure testing, standard setting, and the legal defensibility of high-stakes tests and he has authored several reports for the ETS Visiting Committee. He has served as a consultant to medical board examinations, bar associations, state education agencies, and courts regarding licensure testing, in addition to providing direct testimony in court cases.

Dr. Mehrens earned his doctorate at University of Minnesota in educational psychology.