Education Policy Analysis Archives

Volume 7 Number 4 **February 11, 1999** **ISSN 1068-2341**

A peer-reviewed scholarly electronic journal
Editor: Gene V Glass, College of Education
Arizona State University

Copyright 1999, the **EDUCATION POLICY ANALYSIS ARCHIVES**.
Permission is hereby granted to copy any article
if **EPAA** is credited and copies are not sold.

Articles appearing in **EPAA** are abstracted in the *Current Index to Journals in Education* by the ERIC Clearinghouse on Assessment and Evaluation and are permanently archived in *Resources in Education*.

Less Truth Than Error?
An independent study of the Massachusetts Teacher Tests

Walt Haney, Clarke Fowler, Anne Wheelock, Damian Bebell and Nicole Malec

Ad Hoc Committee to Test the Teacher Test

Center for the Study of Testing, Evaluation and Educational Policy
Boston College

Abstract

The Massachusetts Teacher Tests (MTT), introduced last year, have never been subject to external review as required by the measurement profession's standards and many legal precedents. Neither the Massachusetts Department of Education (DOE) nor the tests' manufacturer have made public information about the exams' reliability (consistency) or validity (meaningfulness). Using data from state and academic reports from the April and July test dates, an ad hoc committee of nationally-known researchers has now been able to make a preliminary assessment of the exams. The committee focused on the Communications and Literacy exam that was required of all prospective teachers regardless of grade level or subject area. The purpose of the analysis was to determine the accuracy of the tests in assessing the reading and writing skills of the test-takers.

Scores on the Massachusetts Teacher Tests of reading and writing are highly unreliable. The tests' margin of error is close to double to triple the range found on well-developed tests. A person retaking the MTT several times could have huge

fluctuations in their scores even if their skill level did not change significantly. In fact, the 9 to 17 point margin of error calculated for the tests represents more than 10 percent of the grading scale (assumed to be 0 to 100). The large margin of error means there is both a high false-pass rate and a high false-failure rate. For example, a person who received a score of 72 on the writing test could have scored an 89 or a 55 simply because of the unreliability of the test. Since adults' reading and writing skills do not change a great deal over several months, this range of scores on the same test should not be possible. While this test is being touted as an accurate assessment of a person's fitness to be a teacher, one would expect the scores to accurately reflect a test-taker's verbal ability level. In addition to the large margin of error, the MTT contain questionable content that make them poor tools for measuring test-takers' reading and writing skills. The content and lack of correlation between the reading and writing scores reduces the meaningfulness, or validity, of the tests. The validity is affected not just by the content, but by a host of factors, such as the conditions under which tests were administered and how they were scored. Interviews with a small sample of test-takers confirmed published reports concerning problems with the content and administration.

If the Commonwealth wants high standards for its teaching force, it should use assessments that meet high professional standards. The current MTT fail this criterion. Results from the April and July administrations of the MTT reveal that these new tests are so unreliable and of such poor validity that they are passing candidates who should fail and failing ones who should pass. Therefore, the ad hoc committee recommends:

1. The Massachusetts Board of Education should immediately suspend the administration of the Massachusetts Teacher Tests.
2. The Commonwealth should convene an independent panel to audit the tests' development, administration, and use.
3. An investigation should be launched to uncover why the state contracted with this test developer even after learning of the company's poor past performance in developing tests of this type.

About the Authors

Walt Haney is a professor in the School of Education and Senior Research Associate in the Center for the Study of Testing, Evaluation and Educational Policy at Boston College. He is former editor of the journal *Educational Measurement: Issues and Practice*, advisor to the committee that developed the 1985 *Standards for Educational and Psychological Testing* and author of numerous articles concerning educational testing and evaluation. He has also served as an expert witness in numerous court cases

concerning testing.

Clarke Fowler is a professor in the Education Department at Salem State College. He has taught both preschool and kindergarten and currently teaches courses in early childhood education.

Anne Wheelock, an independent education policy writer and researcher, works for several national foundations and is the author of *Safe To Be Smart: Building a Culture for Standards-Based Reform in the Middle Grades* (1998).

Damian J. Bebell is a doctoral student at Boston College where he is employed at the Center for the Study of Testing, Evaluation and Educational Research. His research interests include educational philosophy, alternative forms of assessment, and homeschooling.

Electronic mail addresses and phone numbers for the authors are: haney@bc.edu (617-552-4199), clarke.fowler@salem.mass.edu (617-524-4704), wheelock@shore.net (802-254-2796), and bebell@bc.edu.

Center for the Study of Testing, Evaluation and Educational Policy Campion Hall, Boston College, Chestnut Hill, MA 02467

Copyright 1999 by the *Education Policy Analysis Archives*

The World Wide Web address for the *Education Policy Analysis Archives* is http://epaa.asu.edu

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-0211. (602-965-9644). The Book Review Editor is Walter E. Shepherd: shepherd@asu.edu . The Commentary Editor is Casey D. Cobb: casey.cobb@unh.edu .

EPAA Editorial Board

Michael W. Apple
University of Wisconsin

Greg Camilli
Rutgers University

John Covaleskie
Northern Michigan University

Andrew Coulson
a_coulson@msn.com

Alan Davis
University of Colorado, Denver

Sherman Dorn
University of South Florida

Mark E. Fetler
California Commission on Teacher Credentialing

Richard Garlikov
hmwkhelp@scott.net

Thomas F. Green
Syracuse University

Alison I. Griffith
York University

Arlen Gullickson
Western Michigan University

Ernest R. House
University of Colorado

Aimee Howley
Ohio University

William Hunter
University of Calgary

Daniel Kallós
Umeå University

Thomas Mauhs-Pugh
Green Mountain College

William McInerney
Purdue University

Les McLean
University of Toronto

Anne L. Pemberton
apembert@pen.k12.va.us

Richard C. Richardson
Arizona State University

Dennis Sayers
Ann Leavenworth Center
for Accelerated Learning

Michael Scriven
scriven@aol.com

Robert Stonehill
U.S. Department of Education

David D. Williams
Brigham Young University

Craig B. Howley
Appalachia Educational Laboratory

Richard M. Jaeger
University of North
Carolina--Greensboro

Benjamin Levin
University of Manitoba

Dewayne Matthews
Western Interstate Commission for Higher Education

Mary McKeown-Moak
MGT of America (Austin, TX)

Susan Bobbitt Nolen
University of Washington

Hugh G. Petrie
SUNY Buffalo

Anthony G. Rud Jr.
Purdue University

Jay D. Scribner
University of Texas at Austin

Robert E. Stake
University of Illinois--UC

Robert T. Stout
Arizona State University

EPAA Spanish Language Editorial Board

Associate Editor for Spanish Language
Roberto Rodríguez Gómez
Universidad Nacional Autónoma de México

roberto@servidor.unam.mx

Adrián Acosta (México)
Universidad de Guadalajara
adrianacosta@compuserve.com

J. Félix Angulo Rasco (Spain)
Universidad de Cádiz
felix.angulo@uca.es

Teresa Bracho (México)

Centro de Investigación y Docencia Económica-CIDE
bracho dis1.cide.mx

Alejandro Canales (México)
Universidad Nacional Autónoma de México
canalesa@servidor.unam.mx

Ursula Casanova (U.S.A.)
Arizona State University
casanova@asu.edu

José Contreras Domingo
Universitat de Barcelona
Jose.Contreras@doe.d5.ub.es

Erwin Epstein (U.S.A.)
Loyola University of Chicago
Eepstein@luc.edu

Josué González (U.S.A.)
Arizona State University
josue@asu.edu

Rollin Kent (México)
Departamento de Investigación Educativa-DIE/CINVESTAV
rkent@gemtel.com.mx kentr@data.net.mx

María Beatriz Luce (Brazil)
Universidad Federal de Rio Grande do Sul-UFRGS
lucemb@orion.ufrgs.br

Javier Mendoza Rojas (México)
Universidad Nacional Autónoma de México
javiermr@servidor.unam.mx

Marcela Mollis (Argentina)
Universidad de Buenos Aires
mmollis@filo.uba.ar

Humberto Muñoz García (México)
Universidad Nacional Autónoma de México
humberto@servidor.unam.mx

Angel Ignacio Pérez Gómez (Spain)
Universidad de Málaga
aiperez@uma.es

Daniel Schugurensky (Argentina-Canadá)
OISE/UT, Canada
dschugurensky@oise.utoronto.ca

Simon Schwartzman (Brazil)
Fundação Instituto Brasileiro e Geografia e Estatística
simon@openlink.com.br

Jurjo Torres Santomé (Spain)
Universidad de A Coruña
jurjo@udc.es

Carlos Alberto Torres (U.S.A.)
University of California, Los Angeles
torres@gseisucla.edu

Introduction

The Ad Hoc Committee is an independent group, whose formation is described in this report. The drafting and publication of this report have been solely the responsibility of the Committee. Nonetheless, we would like to acknowledge the support and help of numerous people and institutions in preparing this report. These include the Commonwealth Education Deans' Council, Boston College, Bridgewater State College, Elms College, Framingham State College, Lesley College, Salem State College, The University of Massachusetts at Boston, and Westfield State College. In particular, we would like to thank the Center for the Study of Testing and Public Policy, which allowed us to use its address as a temporary mailing address for the Committee. Among individuals who have generously assisted us are Irwin Blumer, Mary Brabeck, Joseph Caruso, John Cawthorne, Richard Clark, Bill Dandridge, Patricia Delaney, Anne Harrison, Virginia Harvey, Catherine Horn, Bailey Jackson, Diane Joyce, Joanne McCourt, Patricia O'Brien, Joan Rasool, Maria Sachs, Bob Schaeffer, Kelly Shasby, Dennis Shirley, and Michael Thomas. Also, we thank Larry Ludlow, Ron Hambleton and Dan Koretz who provided helpful reviews of statistical analyses recounted in this report. The report itself is entirely the responsibility of the authors, and has not been sponsored, funded or endorsed by any institutions or individuals who have aided our inquiry. Reviewers of drafts of this report have been generous in offering comments and suggestions, but naturally not all have agreed with all that is written here.

The Ad Hoc Committee was formed out of concern that important decisions were being based on Massachusetts Teacher Tests (MTT) (Note 1) scores without reasonable evidence on their reliability and validity--a clear violation of professional standards concerning testing. These standards, the widely recognized 1985 *Standards for Educational and Psychological Testing* (Note 2), have been in existence for almost 50 years (in current and previous editions) and have been relied upon in numerous legal proceedings that involved testing in state and federal courts. The very first provision of these standards deals with test validity, requiring that:

> Evidence of validity should be presented for the major types of inferences for which the use of a test is recommended. A rationale should be provided to support the particular mix of evidence presented for the intended uses. (Standard 1.1 p. 13).

Other standards call on test publishers or developers to document the reliability of test scores for each total score, subscore, or combination of scores that is reported (Standard 2.1, p. 29); to clearly describe scales used for reporting scores (Standard 4.1, p. 33); and to document the reliability of classification decisions based on licensure or certification tests (Standard 11.3, p. 65). Moreover, Standard 5.1 requires that: "A technical manual should be made available to prospective test users at the time a test is published or released for operational use" (p. 35).

In contravention of these requirements, MTT results are being used to make decisions about prospective teachers in Massachusetts, and about educational policies in the Commonwealth. Our Committee therefore set out to gather evidence on the technical merits of the MTT. Before we recount what information we have gathered and what has been learned from it, we describe the background to our inquiry.

Background

Massachusetts Commissioner of Education Robert Antonucci reported to the Massachusetts Board of Education on December 15, 1997, that he had selected National Evaluation Systems (NES), a test development company based in Amherst, MA, to develop and administer new teacher certification tests for Massachusetts (see appendix 1 for more details on the chronology of the development of the MTT). We learned of this in January, 1998. One of us (Haney) was sufficiently concerned about this prospect that he faxed to then Associate Commissioner of Education David Driscoll a copy of a federal court decision in which NES had been found to have "violated the minimum requirements for professional test development" when it created teacher certification tests for the state of Alabama. (Note 3) Associate Commissioner Driscoll did not respond directly but we learned through an intermediary that NES had been one of only three contractors to bid on the Request for Responses (RFR) issued on February 24, 1997. In February 1998, the Commonwealth commissioned NES to develop tests of communication and literacy skills and of subject matter knowledge. (Note 4)

At that time, the plan was that there would be two trial administrations of the new tests (in April and July, 1998) and that their scores would not count toward certification of eligible teacher candidates. As described in a study guide and registration bulletin released by the Massachusetts Department of Education (DOE) in January 1998, candidates would be required to achieve qualifying scores in order to be certified only beginning with the third administration in October 1998. But then a change was made: on March 25, 1998, the DOE announced that candidates taking the April and July tests would not qualify automatically; instead, they would have to make passing scores to be provisionally certified.

The first MTT tests were given on April 4, 1998. In the ensuing weeks NES convened scoring panels to help determine where passing scores on the new tests should be set. On June 22, acting on the recommendation of NES, the Massachusetts Board of Education voted to set passing scores one standard error of measurement below those that resulted from NES's analyses of scoring panel reviews. This meant that 44% of the 1,800 prospective teachers who took the April tests would have failed.

Then in the midst of his campaign for election as Governor, Massachusetts Acting Governor Paul Cellucci had Board of Education Chairman John Silber convene a special meeting of the Board to reconsider its vote on qualifying scores. Reversing its decision of less than two weeks earlier, the Board voted on July 2 for higher passing scores; now 59% of the April test-takers would fail. (At this meeting, Interim Commissioner of Education, Frank Haydu, appointed after Antonucci had taken a job in the private sector, resigned. Haydu had recommended the lower qualifying scores because the test was untried and subject to measurement error.)

The results of the first MTT administration were mailed to test-takers in early July. They showed that 70% of examinees passed the reading test, 59% the writing test, and widely varying percentages the 32 subject matters tests. Because candidates had to pass all three tests to pass the MTT overall, the overall passing rate was only 41%.

This overall result, with a majority of candidates failing, led to immense negative publicity. Even before the passing scores were adjusted in July, in a speech to the Greater Boston Chamber of Commerce on June 25, Massachusetts Speaker of the House Tom Finneran commented, "I'll tell you who won't be a teacher. The idiots who flunked

that test and flunked so miserably--and, of course, the idiots who passed them." (As reported in the Boston Herald on 6/26/98, by Darrrell S. Pressley, "Dumb struck: Finneran slams 'idiots' who failed teacher tests.")

The second MTT administration took place on July 11, just days after April test-takers had received their results. In a press release dated July 23, 1998, the DOE said:

> MALDEN--State Commissioner of Education David P. Driscoll today released the results of the April 4 Massachusetts Teacher tests showing the pass and fail rates for prospective teachers by the institution they attended.
>
> The 1795 prospective teachers who took tests in April in communication and literacy and in various subjects entered on their registration sheets information concerning the 56 colleges and universities they attended. That data was then verified by the institutions; 54 checked the data and made corrections where necessary.
>
> Because in some cases there is a very small sample of students taking specific subject tests and a small sample of graduates from some of the institutions, results need to be interpreted carefully. Many schools have fewer than ten students who took parts of the test, so making a broad statistical conclusion in those cases is not sound.

Release of results by institution prompted much public hand-wringing among teacher preparation institutions. A front page story in the Boston Globe on July 24, 1998, for example, reported, "Colleges vow to retool after failures on teacher tests" (O'Brien, 1998).

All this heightened our concern that important decisions were being made on the basis of the hastily-developed MTT before their technical merits had been established. Hence in July, at the suggestion of Clarke Fowler, the three of us (Fowler, Haney, and Wheelock) decided to found the Ad Hoc Committee to Test the Teacher Test. Our initial idea was to ask people who had taken both the MTT and some post-collegiate national examination, such as the Praxis (the successor to the National Teacher Examination or NTE) or the Graduate Record Examination (GRE), to send us copies of their score reports. Comparing the two sets of scores would allow us to test the validity of the new MTT against established tests such as the Praxis or GRE. The principle was simple: if the new MTT were valid and reliable, then scores on the new Massachusetts test should be comparable with those from other tests.

Flyers inviting test-takers to send us copies of their score reports were distributed at many of the testing sites for the July 11 administration of the MTT. As of December 1998, more than 30 individuals had generously provided us with copies of their score reports on both the MTT and some other test (including the Praxis, the Millers Analogies Test, the New York State Liberal Arts and Sciences teacher certification test, and the Texas teacher certification test, the ExCET). Among them, however, there were only twelve cases in which people had taken the MTT and the same other test, leaving us with samples that were too small for reasonable statistical analysis. Nonetheless, many individuals who sent us scores spontaneously offered us comments on the new MTT. Hence we decided to interview those willing to be interviewed. These interviews are summarized in part 4 and appendix 3 of this report.

On August 17, detailed results of the July administration were released to the

institutions of higher education whose students took the tests. Seeing a copy of these results for one institution, which listed students' MTT scores from both April and July, we realized that we could use these data to estimate the reliability of the MTT tests. Adults' basic skills in reading and writing would not change much in a three-month period; people could not cram for the July test since no study guide was available; and, in any case candidates did not learn they would have to retake the test until just days before the second administration when the April results were mailed to examinees. Hence, we set out to acquire data from institutions which had students take the MTT in both April and July. The results of this inquiry are presented in part 3 of this report.

However, before describing the nature and results of this inquiry, we point out that many share our misgivings about the lack of technical documentation on the new MTT. For example, testing experts such as George Madaus of Boston College and Ron Hambleton of the University of Massachusetts at Amherst have publicly commented on the lack of documentation about the new MTT (Madaus, 1998; Hambleton, 1999); and in both press releases and letters, the Commonwealth Education Deans' Council has expressed concern about the lack of evidence on the MTT's validity (Jackson, 1998).

On August 24, 1998 Board of Higher Education Chairman James Carlin called a meeting at Framingham State College for deans and presidents of Massachusetts institutions of higher education to discuss implications of the MTT results. At this meeting, DOE spokesman Alan Safran said: "We're committed to full disclosure about this test," and stated that the Department was willing to release all "non-proprietary" data. The Ad Hoc Committee therefore wrote to Acting Commissioner of Education David Driscoll to request three sets of data from the Massachusetts Teacher Tests that would allow us to analyze their psychometric properties. Specifically we asked for:

1. Item level and total test scores (both raw scores and scaled scores) for all examinees who took the Massachusetts Teacher Tests in April. We seek these data in order to be able to analyze the psychometric properties of items on the April test.
2. Item level and total test scores (both raw scores and scaled scores) for all examinees who took the Massachusetts Teacher Tests in July. We seek these data in order to be able to analyze the psychometric properties of items on the July test.
3. Test scores of all examinees (both raw scores and scaled scores) for all examinees who took the Massachusetts Teacher Tests in both April and July. We seek these data in order to analyze the test-retest reliability of the Massachusetts Teacher Tests. (Ad Hoc Committee letter to Driscoll, September 7, 1998)

Acting Commissioner Driscoll did not reply.

More recently, in a November 22, 1998, commentary article in the Boston Globe, "Good teachers need a good test," Emanuel Mason, the chair of Northeastern University's Department of Counseling Psychology, Rehabilitation, and Special Education wrote:

> The primary problem with the MTT in all its versions is its validity. Validity is the degree to which a test measures what it was designed to measure. . . . If a test does not have validity, it does not measure anything. Further, validity should be demonstrated before a test is used as the basis for decisions on licensing or any other outcome. The MTT's developer has yet

to provide evidence of validity, and the test already has been administered three times.

Ignoring well-established professional standards concerning testing, the DOE and NES have not made available any documentation on the reliability and validity of the MTT. We therefore set out to study the technical merits of the tests, to begin to answer the question Mason and others have raised: How good are the new MTT tests?

The Reliability and Validity of the Massachusetts Teacher Tests

Given the publicity that has surrounded the new tests and the questions that have been raised about their validity and reliability, it is not surprising that Massachusetts officials have sought to defend their merits. For example, in his July 7, 1998, editorial in the New York Times, John Silber wrote that the exams had been "validated by teachers and scholars who prepared it . . . [and] again by the panels of distinguished teachers, administrators and college professors who reviewed the questions for fairness and agreed on minimal passing scores." What this defense does not take into account is that a test cannot be validated simply by having people review test questions.

Test validation refers to the meaning of test *scores* and that meaning depends not just on test content, but also on a host of other factors, such as the conditions under which tests are administered and how they are scored. A simple example illustrates this point. Suppose that we have a test made of 50 three-digit addition problems such as 231 + 458 = ? On its surface, this would seem to be a test of ability to add three-digit numbers. Perhaps so, if given in a math class with 20 or 30 minutes to solve the 50 problems. But suppose the test was sprung with little warning on aspiring accountants as a condition for getting a job, and they were given only five minutes to solve the 50 problems. Under these conditions, the test would obviously measure the ability not just to add three-digit numbers, but also to work fast under pressure. Or suppose that answers above 999 were scored correct only if they included a comma between the hundreds and thousands positions (such that 1,200 would be scored correct, but 1200 would not). If examinees were not told of this scoring rule, this would undermine the validity of the test as a measure of addition skills; the scoring rule would in effect test examinees' adherence to a particular convention for writing numbers greater than 999.

This example is directly relevant to the Massachusetts Teachers Tests, for when candidates signed up to take the April exams, they had been told that these were merely practice tests and results would not count toward certification. But less than two weeks before the examination, the DOE announced that the results *would* count toward certification. Moreover, people taking the MTT in April and July had no access to sample tests or details on how questions (such as exercises in summarization and dictation) would be scored. Hence it is impossible to assess how meaningful the MTT scores are simply by reviewing questions that make up these tests.

The concepts of test validity and reliability

So how does one assess the validity and reliability of test scores? The 1985 test *Standards* says:

> Validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences. (AERA, APA & NCME, 1985, p. 9, emphasis added)

Traditionally, three types of validity evidence have been recognized: content-related validity evidence, criterion-related validity evidence, and construct validity evidence. Content-related validity refers to the "degree to which the sample of items, tasks or

questions on a test are representative of some defined universe or domain of content" (AERA, APA & NCME, 1985, p. 10). As Emanuel Mason pointed out in his November 22, 1998, article in the Boston Globe, this is the only form of validity evidence to which state and NES officials have even referred, and even here they have provided no technical documentation as required by the 1985 test *Standards*.

But, since validation refers to the meaningfulness of test scores, validation must also consider evidence on criterion-related validity and construct validity. Criterion-related evidence "demonstrates that test scores are systematically related to one or more outcome criteria" (AERA, APA & NCME, 1985, p. 11). The validity of college admissions tests is often evaluated in terms of the extent to which scores predict success in college as measured by freshman-year grade-point average (a form of criterion-related validity evidence referred to as predictive validity evidence). Another form of criterion-related validity evidence is concurrent validity. This refers to how scores on one test relate to those on another test intended to measure the same trait, when both tests are taken at about the same time. This is the sort of validity evidence that the Ad Hoc Committee was seeking when we asked test-takers to send us score reports on both the MTT and other tests designed to measure communications skills and/or teaching competence. As recounted above, we have not been able to acquire enough data to allow us to undertake a concurrent validity study.

Construct validity is an over-arching concept referring to evidence that test scores represent "a measure of the psychological characteristic of interest" (AERA, APA & NCME, 1985, p. 9):

> The process of compiling construct-related evidence for test validity starts with the process of test development and continues until the pattern of empirical relationships between test scores and other variables clearly indicates the meaning of the test score. Especially when multiple measures of a construct are available -- as in many practical testing applications -- validating inferences about a construct also requires paying careful attention to aspects of measurement such as test format, administration conditions, or language level, that may affect test meaning and interpretation materially.
>
> Evidence for construct interpretation of a test may be obtained from a variety of sources. Intercorrelations among items may be used to support the assertion that a test measures primarily a single construct. Substantial relationships of a test to other measures that are purportedly of the same construct and the weaknesses of relationships to measures that are purportedly of different constructs support the identification of constructs and the differences among them. Relationships among different methods of measurement and among various nontest variables similarly sharpen and elaborate the meaning and interpretation of constructs.
>
> Another line of evidence derives from analyses of individual responses. Questioning test takers about their performance strategies or responses to particular items or asking raters about their reasons for their ratings can yield hypotheses that enrich the definition of a construct. (AERA, APA & NCME, 1985, p. 10, emphasis added).

A test that is valid must be reliable. "Reliability refers to the degree to which test scores are free from measurement error" (AERA, APA & NCME, 1985, p. 19). As a basic textbook on testing points out, "The ceiling for possible validity of a test is set by its reliability" (Thorndike & Hagen , 1977, p. 87). In other words, if a test does not measure something reliably, it cannot be a valid measure of anything.

The reliability of the Massachusetts Teacher Tests

Absent sufficient data to assess the concurrent validity of the MTT, we decided to inquire into their reliability. Specifically, once we realized that relevant data might be available to us, we sought to examine the reliability of the scores on the April and July administrations. Comparing scores on these two administrations of the MTT might be thought of as a study of either test-retest or alternate-forms reliability. Test- retest reliability refers to the consistency of scores on two administrations of a test. Alternate-forms reliability refers to the consistency of scores on two different forms or versions of the same test. As we do not know to what extent the July MTT tests used identical questions as the April MTT tests, it is unclear whether our study should be termed a test-retest or alternate-forms study. But the essential idea is quite simple. It is to compare the scores on the MTT tests for people who took them in both April and July. Adults' scores on basic skills tests of reading and writing tests should not change much over three months, and since no study guides were available, examinees could hardly have crammed for the second administration. So if the MTT tests are reasonably reliable, we would expect individuals' scores on these two administrations to be similar; if they are unreliable, we would expect the scores to vary widely.

We should acknowledge that there are several other ways in which test reliability might be estimated, such as internal consistency (indicating how much item results on one test administration tend to cohere.) (Note 5) But, the goal of certification tests such as the MTT is to estimate not simply examinees' competence on one test-taking occasion, but their competence in general. Alternate-forms reliability is thus more appropriate for assessing reliability. As Thorndike and Hagen (1977, p. 79) point out, "evidence based on equivalent test forms should usually be given the most weight in evaluating the reliability of a test."

After the July administration of the MTT, the Department of Education distributed lists of results for individual test-takers to institutions of higher education in the Commonwealth. When we realized that these lists contained individuals' scores for both the April and July tests, we decided to try to gather enough data to undertake a test-retest or alternate- forms reliability study. We contacted some institutions individually via phone, and some through the Commonwealth Education Deans' Council. Three institutions with large numbers of students retaking the MTT tests in July were contacted via letter, which read in part:

> We are seeking your cooperation in affording us access to data that will allow us to analyze some of the psychometric properties of the Massachusetts Teacher Tests.
>
> In particular we seek access to scores of examinees who took Massachusetts Teacher Tests in April and again in July. Access to these data will allow us to analyze the test-retest reliability of the Massachusetts Teacher Tests. Your institution received a computer printout labeled "Institution Roster

Report By Test: Verified Institutional Affiliation" for test date July 11, 1998. We would like to receive either a copy of this computer printout or a data file containing relevant data. To preserve the confidentially of examinees we seek these data with the names and SSN's of candidates removed. (Letter to institutional representatives, September 22, 1998).

As of mid-December we had received MTT data from eight institutions, namely Boston College, Bridgewater State College, Elms College, Framingham State College, Lesley College, Salem State College, UMass Boston and Westfield State College. Five of these institutions are public and three are private. In both April and July, students from over 50 different institutions took the MTT. Eight out of 50 represents only 16% of the institutions that had students taking the MTT, but since these eight represent some of the largest teacher training institutions in the Commonwealth, they account for close to one third of the candidates who took the MTT in April.

Altogether we collected data on 219 people who took the MTT tests in both April and July, though not all 219 took the reading, writing and subject matter portions of the MTT on both occasions. (Note 6) One of the first things we noted about the April and July MTT scores is that some of the score changes seemed truly bizarre. (Note 7) For example, one individual was reported to have scored 36 on the reading test in April and 75 on the April writing test, but to have scored 89 on the reading test in July and 17 on the writing test. In another case, an individual was reported to have scored 56 on the writing test in April and then got an 11 in July. Such huge score changes seemed so unlikely that we inquired into the accuracy of the reported scores. In both of these cases, the scores reported were verified by institutional representatives. In the first case we were told that the individual had not been prepared for the reading test in April, and that the dramatic increase from 36 in April to 89 in July was explained by the fact that the test taker had known that the latter counted toward certification. Why did the writing score plummet from 75 to 17 while the reading score increased from 36 to 89? According to the institutional representative, this happened because the test-taker started taking the July writing test, but then remembered that because she had scored more than 70 in April, she did not have to take the writing test again in July. Hence she simply stopped answering questions. Nonetheless the July score of 17 was reported to the institution as a failure. In the second case, in which writing scores dropped from 56 in April to 11 in July, the institutional representative verified the accuracy of the scores. She had no explanation for the dramatic score decrease, but added that the individual who had received these scores had left Massachusetts to take a teaching job in Arizona.

Table 1 presents the summary statistics for the 219 cases of April and July MTT test-takers for which we have data.

Table 1: Summary Descriptive Statistics on April-July MTT Test-Takers

	Reading 4/98	Writing 4/98	Reading 7/98	Writing 7/98
Count	215	218	130	173
Mean	65.2	63.1	69.4	70.7
Median	66	63.5	70	71

Standard deviation	14.7	10.3	15.2	11.8
Minimum	3	36	21	11
Maximum	93	87	96	96

These data suggest that this sample is not unlike the April MTT test takers in general. On average, they fell below the passing score of 70 on both the reading and writing tests. Initially, these results would appear to make the MTT results seem reasonably reliable. Among repeat test-takers, the average reading scores increased from 65.2 to 69.4, and the average writing test scores from 63.1 to 70.7, apparently modest changes. But note the differences in the count of people in this sample who took the April and July tests. While more than 200 took both reading and writing tests in April, fewer than 180 took the tests in July. This reflects the fact that test-takers had to retake tests in July only if they had scored less than 70 on either the reading or writing tests.

Hence, in order to assess the reliability of the MTT tests, we need to examine the correlations between April and July tests for examinees who took the same portions of the MTT tests on both occasions. Table 2 shows the intercorrelations of reading and writing test scores for people who took both tests. For those who took the reading test in April and July, the correlation of scores was 0.29; for writing, 0.37.

Table 2: Intercorrelations of April and July MTT Scores

	Reading 4/98	Writing 4/98	Reading 7/98	Writing 7/98
Reading 4/98	1.00 (215)	0.07 (215)	**0.29** **(127)**	0.24 (169)
Writing 4/98		1.00 (218)	0.47 (129)	**0.37** **(172)**
Reading 7/98			1.00 (130)	0.06 (94)
Writing 7/98				1.00 (173)

Note: Sample sizes shown in parentheses, test-retest correlations in bold.

These test-retest intercorrelations are extraordinarily low. Correlation coefficients can range from -1.00 to +1.00. Test-retest correlation coefficients for well-developed standardized tests typically range between 0.80 and 0.90. For example, test-retest correlations for the SAT have been reported to range between 0.86 and 0.90 (Donlon,

1984, p. 54). Similarly, Thorndike & Hagen (1977, p. 321) report alternate-form reliability coefficients in the range of .85 to .95 for the Stanford Binet. In contrast, the scores of examinees who took MTT reading tests in April and July correlated only 0.29, those of examinees who took the writing test in April and July correlated 0.37.

To provide a more detailed picture of the relationship between April and July scores, Figures 1 and 2 show scatterplots of test scores for individuals in our sample who took the reading and of writing tests, respectively, on both occasions. Several patterns are apparent in comparing these figures. First, note that the "scatter" in reading test scores is greater than that in writing scores. This simply reflects the findings shown in Table 2 above; namely, that the correlation between reading scores in April and July (0.29) was smaller than that for writing test scores (0.37).

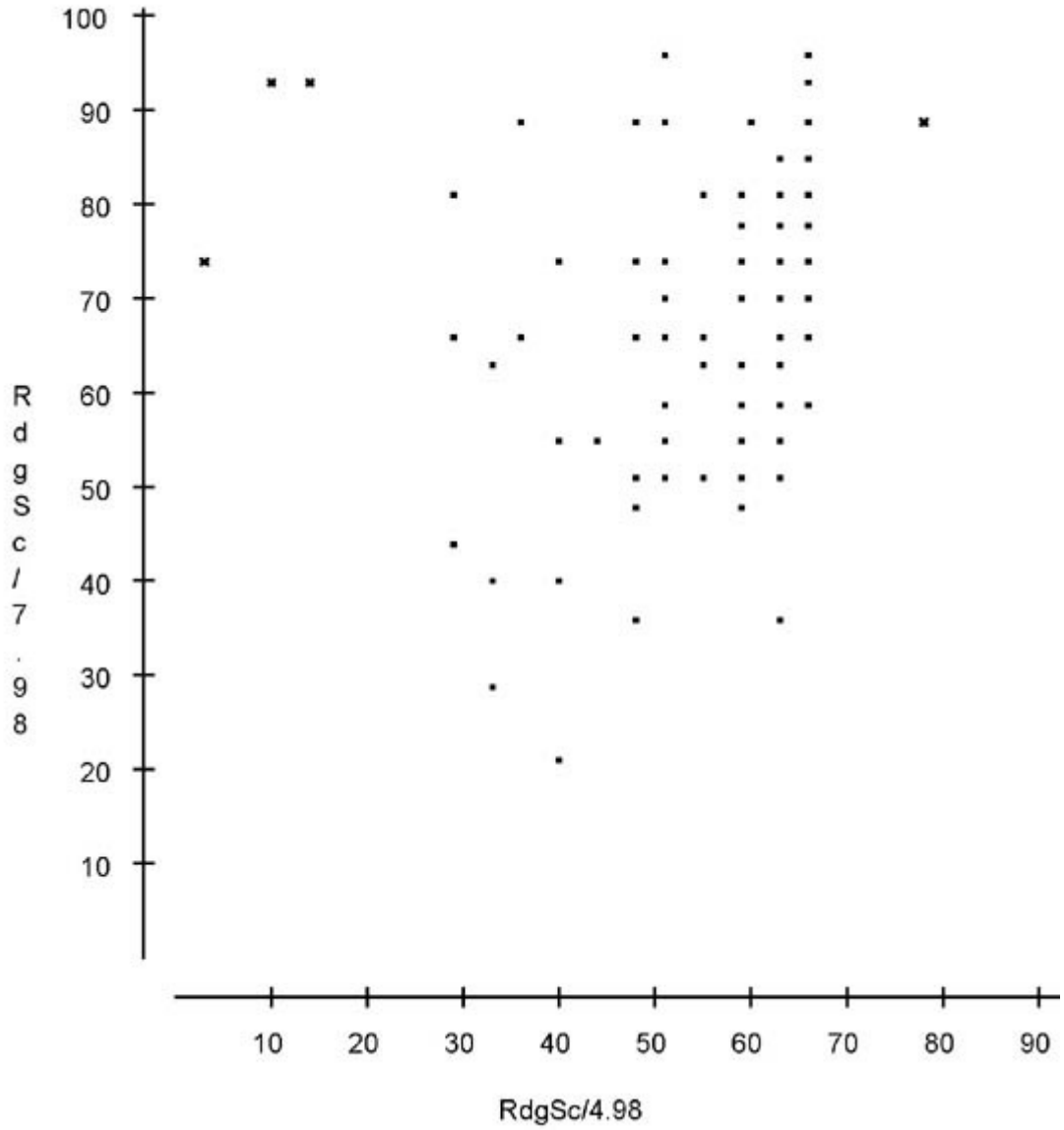

Figure 1. Scatterplot of April and July MTT Reading Test Scores

Note also how widely retest scores vary among people who had approximately the same test scores in April. For example, in Figure 1, among examinees who had scores of about 60

on the reading test in April, retest scores in July ranged from less than 40 to about 90. And, as is apparent in Figure 2, among test-takers who scored in the 65 to 69 range in April, retest scores range from about 50 to 90.

These figures also illustrate some of the huge score changes that initially caught our attention. These cases, often called "outliers" in data analysis, are marked with x's in Figures 1 and 2. In Figure 1, for example, note the three cases in the upper left corner. In all three cases, examinees had scores of less than 20 on the reading test in April but more than 70 in July, increases of more than 3 standard deviations. And in Figure 2, note the case in the lower right corner, representing someone who had a score of 75 in April, but a score of 17 in July. This is the case mentioned previously that was so bizarre that we asked the institutional representative to verify the accuracy of the data--the case of the test-taker who, remembering she did not have to take the writing test again, simply stopped answering questions . (Note 8) The other clear "outlier" in Figure 2 is lowest x on the figure, representing someone who had a score of 56 on the writing test in April, but 11 in July.

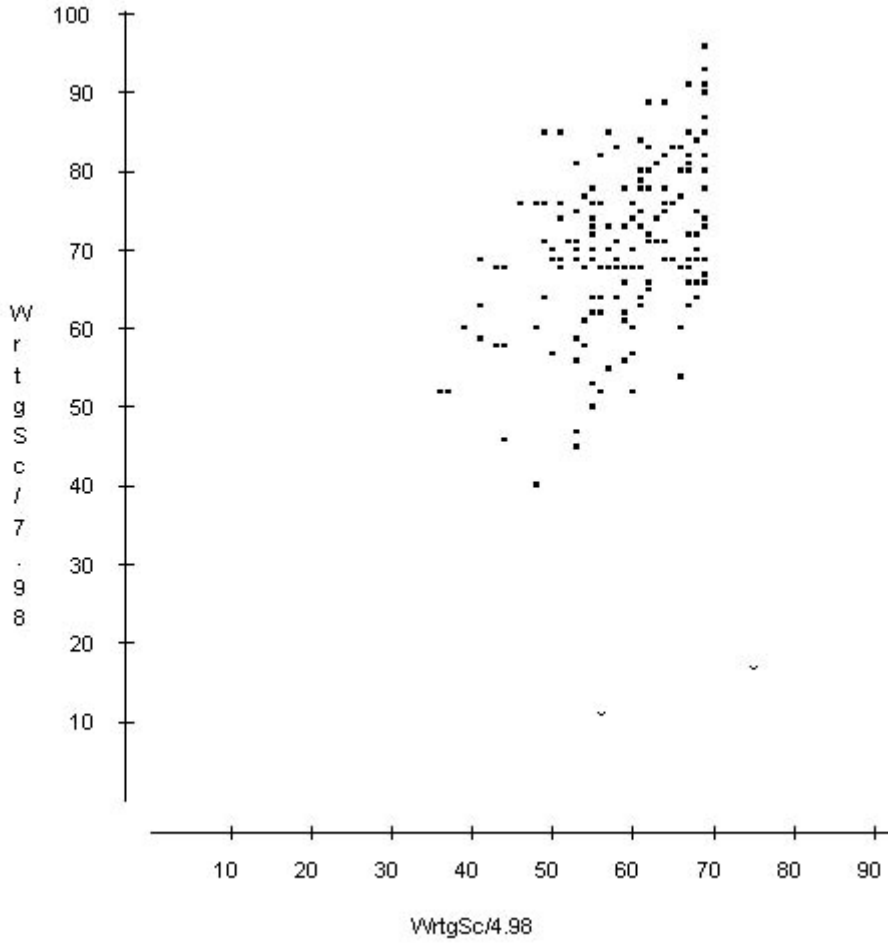

Figure 2. Scatterplot of April and July MTT Writing Test Scores

We have checked these "outlier" cases and all are accurate in terms of scores reported to institutions. Nonetheless, as a more conservative examination of the test-retest reliability, we recalculated the test-retest correlations after deleting the outliers. We refer to these groups, after deleting outliers, as our trimmed samples. Specifically, after deleting the four unusual cases marked in Figure 1 with x's, the test-retest correlation for the reading test rose to 0.49.

Similarly, after deleting the two outlying cases shown in Figure 2, the test-retest correlation for the writing test increased to 0.48.

This brings us to one other feature apparent in Figures 1 and 2, and also to a possible explanation for the remarkably low test-retest correlations shown in Table 1. Note that in both Figures 1 and 2, there is only one case for which retest data are available for an examinee who scored 70 or above in April. This is because people who scored 70 or above "passed" the tests and did not have to retake them in order to be provisionally certified. With this one exception, our test-retest data for the MTT are for people who scored below 70 on the April tests. This leads to one possible explanation for the unusually low test-retest correlations, namely attenuation of observed correlation coefficients due to restriction of range. This concept is easy to explain with an example. People's height tends to be correlated with their weight. Tall people tend to weigh more than short people. Thus, we would find a positive correlation between the heights and weights of adults in general. But suppose that we consider a sample of people who are all exactly five feet tall. If we examine the correlation between their height and weight, we will find a zero correlation for the simple reason that they are all of the same height. By focusing on people who are exactly five feet tall, we have restricted the range on this variable; hence, the observed correlation between height and weight for this sample has been reduced or attenuated. This is what is meant by attenuation due to restriction of range. If we restrict the range of a variable, the observed correlation between this variable and another will be attenuated, as compared to the correlation that likely would be observed if the range on the variable were not restricted.

Hence, before concluding that the MTT reading and writing tests are unreliable, we need to consider the possibility that attenuation due to restriction of range, with most of test-retest data available only for examinees who scored less than 70 on the April tests, may have led to the low test-retest correlations shown in Table 2. Fortunately, the phenomenon of attenuation of correlation coefficients due to restriction of range has been widely recognized in the testing and measurement literature. Formulas and tables are available to allow estimation of "unattenuated" correlation coefficients when restriction of range is taken into account (slightly different formulas are available, for example, in Lord & Novick, 1968; Cronbach, 1971; and Linn, 1982).

Lord & Novick (1968) present an extended discussion of attenuation due to restriction of range and tables showing how observed correlations can be corrected for attenuation. If we assume that the relationship between two variables is linear and that the conditional variance of one does not depend on the particular value of the other (the assumption of homoscedasticity), then the following table shows the corrections for observed correlations when the percentage of the sample is restricted to the top (or bottom) 60%, 50%, 40% and 30% of the entire population. As shown in Table 2 above, we found that the observed correlation between the April and July MTT reading tests was 0.29. However, 70% of examinees passed the April reading test, so the range of examinees who had to take the July reading test was "restricted" to only the bottom 30% of the population of April examinees. Table 3 indicates that a correlation of 0.30 observed when range is restricted to 30% of a population would be corrected to 0.519 for the whole population. Similarly, we observed a correlation of 0.37 between scores on the April and July writing test, but since about 60% of examinees passed the April writing test, the group retaking the July writing test was restricted to about 40% of the population. Table 3 indicates that an observed correlation of 0.40 in a sample restricted to 40% of a population would be corrected to 0.616 for the entire population. For the trimmed samples, the observed correlations of 0.49 and 0.48, would be corrected to about 0.74 and 0.72, again presuming that only the bottom 30% retook the reading test and the bottom 40% retook the writing test.

Table 3: Corrections for Attenuation Due to Restriction of Range						
Normal deviate z	Percent selected in restricted sample	Standard devia-tion in selected group	Ratio of SD in unselected to SD in selected groups (K)	Observed correlation of 0.30 in restricted sample corrected to	Observed correlation of 0.40 in restricted sample corrected to	Observed correlation of 0.50 in restricted sample corrected to
-0.25	59.9	0.65	1.54	0.436	0.558	0.644
0	50	0.6	1.64	0.458	0.582	0.688
0.25	40.1	0.56	1.79	0.491	0.616	0.719
0.5	30.8	0.52	1.93	0.519	0.644	0.744

Source: Adapted from Lord & Novick, 1968, pp. 140-142.

To verify these corrections for attenuation due to restriction of range, we conducted simulation analyses to address questions such as the following. If the test-retest correlation among a group of test takers was 0.50, what would be the correlation observed if only the bottom 30% on the initial test were considered? We do not attempt to present all of the results of these simulations here, but instead, in Figure 3, present the results of one iteration of the data simulations aimed at addressing the following question. If there were a test- retest correlation between test 1 (t1) and re-test (t2) of 0.50, what would be the observed correlation between test and re-test scores if attention were restricted to only the bottom 30% on the initial test (t1). What our results show is that if there were a test-retest correlation of 0.50 among the entire population of test-takers, restricting attention to only the bottom 30% of test-takers on the initial test (t1) would reduce (or attenuate) the observed correlation to about 0.30. These results confirm the theoretical results reported above. Given that we observed a test-retest correlation of about 0.30 in the 30-40% of examinees who had to retake the MTT, our estimate of the test-retest correlation for the MTT, if all examinees had retaken the tests, is about 0.50.

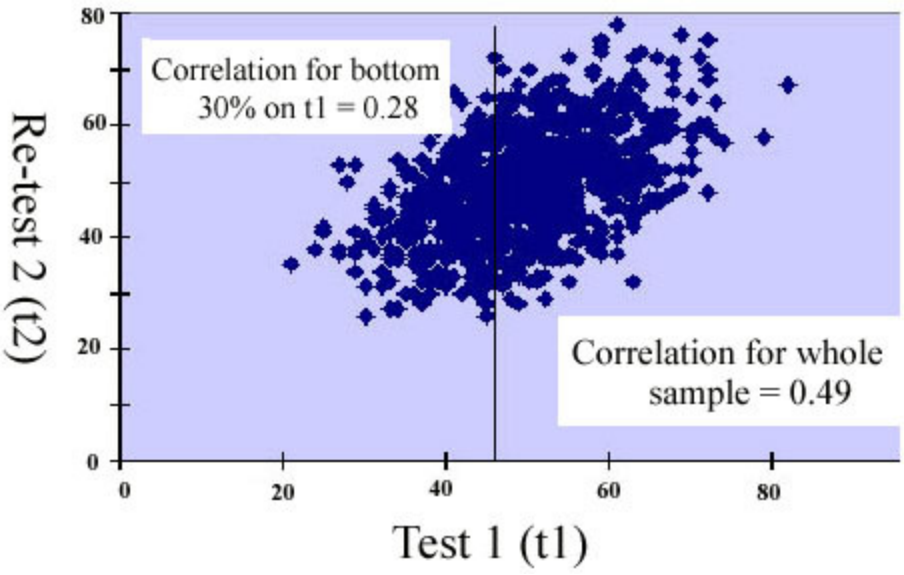

Figure 3. Example of Simulation Results

Note: Results shown here are for a sample of 1000

Test-retest correlations in the range of 0.50 (or even 0.70) are unusually low. In comparison, as previously mentioned, test-retest correlations for the SAT have routinely been found to be in the range of 0.85 to 0.90 (Donlon, 1984, p. 54). There are several ways of illustrating the implications of test-retest reliability being as low as 0.50. One way of interpreting a test-retest reliability coefficient rtt is as the ratio of signal to "noise plus signal," or as the ratio of true score variance to observed score variance.

$\mathbf{r}_{tt}$ = signal / (signal + noise) = true score variance / observed score variance

Since observed score variance is composed of true score variance plus error score variance (see Anastasi, 1976, pp. 120-22, or many other textbooks on testing, for more detailed explanations), this equation can also be expressed as

$\mathbf{r}_{tt}$ = (true score variance) / (true score variance + error score variance)

Thus, it is easy to see that when a test-retest reliability coefficient $\mathbf{r}_{tt}$ is as low as 0.50, observed scores are composed of as much error score variance as of true score variance. Thus a test-retest correlation of 0.50 indicates that MTT scores contain as much error as true score variance. Even a test-retest correlation of 0.70 indicates that MTT scores are composed of 30 percent error variance.

A second way of showing the meaning of a test-retest reliability coefficient $\mathbf{r}_{tt}$ is to use it to calculate the standard error of measurement, as follows:

$$e = s_t \sqrt{1 - r_{tt}}$$

(Thorndike & Hagen, 1977, p. 85; Anastasi, 1976, p. 128)

where:

e = standard error of measurement
$\mathbf{s}_t$ = standard deviation of test scores, and
$\mathbf{r}_{tt}$ = test-retest reliability coefficient.

As shown in Table 2, in our test-retest sample, we found the standard deviations of reading and writing test scores to be about 15 and 11 points respectively. However, these observed standard deviations were based on the restricted sample of retest examinees (with only 30% of April examinees having to retake the reading test and 40% the writing test), so we need to find a way of estimating the standard deviations of MTT test scores for the entire population of test takers.

As we have pointed out, even after the MTT have been administered four times, over a period of a year, no technical report on these new tests has been issued. Hence, we must rely on data shared with us by cooperating institutions to estimate the standard deviations of MTT reading and writing test scores among the entire population of examinees. We have available two different avenues for pursuing this end; usingtheoretical adjustments of data on our test-retest sample and using data institutions shared with us on all their students who took the MTT in April and July.

page 1 | introduction | background | reliability & validity | interviews | conclusions | references

Using the theoretical approach (and the data shown in Table 3 above), we can multiply the restricted sample standard deviations by 1.93 and 1.79 to estimate the standard deviations in the full population of April examinees. Since 15 x 1.93 = 28.95, and 11 x 1.79 = 19.69, we may use these figures as one set of estimates of the standard deviations of the MTT reading and writing tests. A second approach is to examine the standard deviations of the April tests for the institutions which gave us data on all of their April test-takers. We found that that the within-institution standard deviations to be as high as 19 points for the April reading test and 16 for the April writing test.

Hence, as summary estimates of the standard deviations of the April tests, we averaged these two estimates, which yielded 24 [(29 +19)/2] and 18 [(16 +20)/2] as ballpark estimates of the standard deviations of the April tests for the full population of test takers. Then we estimate standard error of measurement for the MTT reading and writing tests as follows:

$$e_r = 24\sqrt{1 - 0.519} = 24\sqrt{0.481} = 24 \text{ x } 0.694 = 16.6$$

$$e_w = 18\sqrt{1 - 0.616} = 18\sqrt{0.384} = 18 \text{ x } 0.62 = 11.2$$

Even if we use the more conservative estimations of test-retest correlations, based on the trimmed samples (that is, with outliers deleted) and adjusting for attenuation dues to restriction of range, namely 0.74 and 0.72 for the MTT reading and writing tests respectively, these would still imply standard error of measurement of 12.2 and 9.5. In other words, our results suggest that the standard errors of measurement in the April MTT Reading and Writing tests were about 17 and 11 points respectively (or at best 12 and 9). While neither the Massachusetts DOE nor NES has yet released any technical information on the scaling of the MTT, we have found MTT scores to range from near zero to almost 100. If indeed the scores for the MTT are on a 100 point scale, this means that standard errors of measurement of 9 and 17 points represent some 9% to 17% of the entire score range. This means that

examinees scoring in the range of 50 to 69 may easily have "failed" the MTT simply because of measurement error, and, conversely, ones scoring in the range of 70 to 90 may well have "passed" simply because of the large degree of measurement error in the MTT tests.

These errors of measurement on the MTT may be compared with the standard error of measurement on well-known tests for which technical documentation is available. The SAT (originally, the Scholastic Aptitude Test, briefly renamed the Scholastic Assessment Test, and now just the SAT) is reported on a scale that ranges from 200 to 800, or 600 points. The standard errors of measurement of the SAT verbal and quantitative scores have been reported to be in the range of 29-34 points (Donlon, 1984, pp. 33-34), or 4.4 to 5.7% of the score range. The standard errors of measurement for the Graduate Record Examination have been reported to be 33, 38 and 36 points for the GRE Verbal, Quantitative and Analytical subtests respectively (Conrad, Trismen & Miller, 1977, p. 19). Since these scores are reported on scales ranging from 600 to 670 points, these standard errors of measurement are all less than 6% of scale range. If the standard errors of measurement of MTT scores are 9 to 17 points, as we have estimated, representing 9% and 17% of the MTT score range, this means that MTT scores have almost double to triple the degree of error as the SAT and the GRE (as estimated by SEM relative to scaled score range).

The reliability of classifications based on the MTT

This brings us to a point mentioned in the introduction of this report. As the 1985 *Standards for Educational and Psychological Testing* point out, for licensure or certification tests on which people are rated as passing or failing, it is important to provide data not just on test scores, but also on the reliability of classification decisions based on those scores (Standard 11.3, p. 65). As Ron Hambleton has pointed out, "there were serious problems with the setting of passing scores on the reading literacy, writing and subject matter [MTT] tests":

1. A detailed description of what it means to be qualified was not developed;
2. Panelists who set passing scores did not have an opportunity to discuss their recommendations with each other prior to finalizing their recommendations to the Board;
3. Technical data arising from the process of setting passing scores was not presented to the Board for their consideration. (Hambleton, 1999, pp. 20-21)

Nonetheless, data available from the April and July administrations of the MTT allow us to examine the reliability of pass/fail classifications based on the MTT reading and writing tests. In a report entitled "Massachusetts Teacher Tests. Summary of Institution results for Second-Time Test Takers. Test Date: July 11, 1998. Test Summary," the Massachusetts DOE released data shedding direct evidence on this point. The report lists the number, and percent passing, of examinees who retook the MTT on July 11, 1998. Results are reported only for institutions that had more than four candidates. Hence this table showed reading test results for only 18 institutions and writing test results for 23 institutions. These data are shown in Table 4 below.

Table 4: Passing Rates of Second-Time Test Takers on July 11, 1998, MTT Tests

Institution	Reading	Reading %	Writing	Writing %

	N	pass	N	pass
American International College	10	70.0	13	30.8
Anna Maria College	5	20.0	5	40.0
Assumption College			6	50.0
Boston College	7	85.7	5	80.0
Bridgewater State College	41	58.5	59	55.2
Curry College	6	33.3	6	16.7
Fitchburg State College	25	56.0	31	45.7
Framingham State College	16	31.3	19	52.6
Lesley College	12	66.7	14	64.3
Mass. College of Liberal Arts	6	33.3	10	50.0
Merrimack College	5	40.0	9	66.7
Salem State College	17	76.5	26	76.9
Simmons College			6	100.0
Springfield College	22	36.4	27	48.1
Stonehill College	10	90.0	13	53.8
Univ. of Massachusetts/ Amherst	16	68.8	20	35.0
Univ. of Massachusetts/ Boston			6	50.0
Univ. of Massachusetts/ Dartmouth			6	50.0
Univ. of Massachusetts/ Lowell	5	60.0	13	76.9
Westfield State College	27	48.1	32	34.4
Wheelock College	6	50.0	12	50.0
Worcester State College	11	72.7	15	60.0
Unaffiliated candidates	35	60.0	47	44.7
Mean		55.6		53.6
Median		58.5		50.0

Source: Adapted from Mass DOE, "Massachusetts Teacher Tests. Summary of Institution results for Second-Time Test Takers. Test Date: July 11, 1998. Test Summary."; (Available at http://www.doe.mass.edu/teachertest/)

As the data in Table 4 indicate, the mean pass rate (unweighted average across institutions for which data were reported) among second-time MTT test-takers was over 50% on both the reading and writing tests. Though we do not show weighted results in Table 4, these data indicate that 160 of 282 or 57% of examinees taking the reading test for the second time passed, and 207 of 400 or 52% of those taking the writing test passed. This indicates that the misclassification rate among those who "failed" the April tests was over 50% on both the reading test and the writing test. This seems extraordinarily high given that adults' basic skills in reading and writing are unlikely to change much over a three month-period (and as previously mentioned, candidates could not cram for the July test). Note, too, that the misclassification rate was higher on the reading test than on the writing test-- exactly what would be predicted from the results of our reliability analysis, which showed the reading test to be less reliable than the writing test.

page 1 | introduction | background | reliability & validity | interviews | conclusions | references

What these results do not show, of course, is the rate of misclassification of those who passed the MTT tests. Nonetheless, it seems certain, given the results of our analyses, that a substantial proportion of those who "passed" the MTT reading and writing tests by 10 to 20 points did so simply because of test unreliability. People who scored above 69 on the MTT reading and writing tests, and thus "passed" these tests in April did not have to retake them in July. Hence we have no direct way of estimating the false "pass" rate. But to get a rough idea of this kind of misclassification we examined the percentage of April examinees scoring in the 50-69 point range whose scores decreased. We found that some 4% to 10% of first-time test-takers in these score ranges had decreased scores upon retest. Thus, a very conservative estimate of the percentage of April examinees who "passed" simply because of measurement error would be 2% to 5%.

Table 5: Passing Rates of Second-Time Test Takers Reported by DOE compared with Retest Sample

	Reported by		Mass DOE		Ad Hoc		Retest Sample	
Institution	Rdg N	Rdg % pass	Wrtg N	Wrtg % pass	Rdg N	Rdg % pass	Wrtg N	Wrtg % pass
Boston College	7	85.7	5	80.0	6	83.3	5	80.0
Bridgewater State College	41	58.5	59	55.2	42	59.5	60	56.7
Framingham State College	16	31.3	19	52.6	16	31.3	19	52.6
Lesley College	12	66.7	14	64.3	12	66.7	14	64.3
Salem State College	17	76.5	26	76.9	21	66.7	34	79.4
U. Mass/Boston			6	50.0			6	50.0
Westfield State College	27	48.1	32	34.4	27	48.1	32	34.4

Source: Adapted from Mass DOE, "Massachusetts Teacher Tests. Summary of Institution results for Second-Time Test Takers. Test Date: July 11, 1998. Test Summary". (Available at http://www.doe.mass.edu/teachertest/)

The data summarized in Table 4 also allowed us to check the findings from our test-retest sample against passing rates reported by the DOE that are summarized in Table 5. This table presents the passing rates reported by the DOE with those apparent in our test- retest sample. Note first that this table shows no results for Elms College; the DOE did not report any results for this institution because it had fewer than five second-time test takers. For four of the remaining seven institutions, the sample sizes (Ns) and passing rates reported by the DOE are exactly the same as those in our test-retest sample. For the remaining three institutions there are slight differences between results reported by the DOE and those apparent in our test-retest sample. For Boston College, the DOE reported seven second-time takers for the reading test, whereas we counted only six in our test-retest sample. We have examined the data for Boston College in detail and suspect that this discrepancy arises from an unusual case in which one student took the MTT in April, and had an April writing test score reported, but had no April reading test score reported in results transmitted to Boston College. Thus, apparently this individual was counted in the DOE results as a second-time test-taker, but was not included in our test-retest sample because no reading test score was reported for April. The other two institutions for which there are slight discrepancies are Bridgewater and Salem. In both cases, the Ns for the test-retest sample are slightly higher than the Ns reported in DOE results. These differences apparently derive from the fact that the DOE results are reported only for individuals whose institutional affiliation was verified by the institution by a particular date. The reason for the slightly larger Ns for Bridgewater and Salem is that the data provided to us apparently included a small number of cases that were treated as unaffiliated examinees by the DOE.

Indeed, the DOE's policy of institutional affiliation of MTT test takers seems to be of doubtful merit and of changing meaning. In reporting institutional results for the April, July, and October administrations of the MTT, the DOE has offered a number of slightly different "Interpretive cautions and notes." But in each instance, the first has read as follows:

> 1. Information regarding candidate institutional affiliation was obtained from candidates as self-reported information on the registration form during the test registration process. This information was forwarded to institutions of higher education, which were provided with an opportunity to verify the candidates' institutional affiliation. The institutions were informed that if they did not respond to the verification request as explained, the data to be included in their results would be based on candidate-reported institutional affiliation.

The institutional results for the April administration of the MTT were released under a memo from David Driscoll dated July 21, 1998. Together with institutional results for the July and October 1998 administrations, they are also available on the DOE web site. When we examined results for April and July, we noticed that there was a sharp increase in the number of "unaffiliated" candidates, that is, ones whose affiliation was with institutions outside Massachusetts or was not verified by the institutions concerned. Hence, we used the data available from the DOE web site to calculate the percentages of test-takers at each administration who were listed as "unaffiliated." As the results in Table 6 show, between April and July there was a fourfold increase in the percentage of test-takers listed as unaffiliated.

Table 6: First-time Test-Takers Listed as Unaffiliated

MTT Test	APRIL	JULY	OCTOBER
Reading	227 / 1794= 12.7%	891 / 1702= 52.4%	778 / 1533= 50.8%
Writing	227 / 1808= 12.5%	898 / 1707= 52.6%	783 / 1544= 50.7%

Sources:
http://www.doe.mass.edu/teachertest/7981st.html
http://www.doe.mass.edu/teachertest/summary498.html_
http://www.doe.mass.edu/teachertest/1098inst/1test.html
(Data summarized in Table 6 were downloaded 1/5/99)

One possible explanation for this sharp increase is that students enrolled in out-of- state colleges were more readily able to come to Massachusetts to take the MTT in July than they were in April. This is not confirmed, however, by the fact the proportion of first-time test-takers listed as unaffiliated remained very high, more than 50%, in the October administration. Thus, what appears to have happened is that Massachusetts institutions of higher education with teacher preparation programs changed the manner in which they verified the affiliation of students after the first administration of the MTT. Indeed, on September 26, 1998, an article in the Boston Globe, "BU test to screen teacher hopefuls, disclaim failures," reported that Boston University had instituted a policy of not verifying students' affiliation with BU unless they had passed a literacy screening test before taking the MTT (Zernike, 1998).

What the BU policy and the fourfold increase in the percentages of unaffiliated candidates indicate, however, is that even if the MTT test scores were reliable and valid, the DOE's practice of publishing "institutional" results may be highly misleading unless some better and uniform methods of "verifying" candidates' affiliation with institutions is developed. And even if that problem were solved, the ranking of schools based on student test scores is of doubtful merit.

Ranking schools and school districts (and even states and countries) on student test results seems to be increasingly popular with the media in recent years. This is both unfair and ineffective in improving education. It is unfair for the simple reason that judging the effectiveness of educational institutions should be based not on end-of-school test scores, but on "value-added" as a result of experience in the school. If a value-added perspective is not adopted, then highly selective institutions (such as Harvard, Boston College, and Boston University among teacher preparation institutions in Massachusetts) may come out looking good in such rankings, simply because they only admit students who are good test-takers to begin with, not because of how much students learn while attending them.

And ranking of schools based on student test scores can be both unfair and ineffective unless attention is paid to not just test scores as "outcomes," but also to the educational processes that produced those outcomes. In recent years, for example, numerous cases have been revealed in which schools cheated on tests in order to make their rankings look better. And even absent such manipulation of test scores, lack of attention to processes provides little leverage for improvement (Haney & Raczek, 1993).

The content validity of the Massachusetts Teacher Tests

Validity refers to the appropriateness and meaningfulness of inferences drawn from test scores. As explained previously, three types of evidence have been recognized in this connection; content-, criterion- and construct-related validity evidence. The Ad Hoc

Committee originally intended to gather one form of criterion-related evidence, namely concurrent validity evidence that would compare scores on the MTT with those on well-established tests for college graduates. Having failed to gather enough comparable scores to allow reasonable statistical analysis, we undertook the reliability studies described above. Our findings that the MTT reading and writing tests are unreliable and have led to remarkably high rates of misclassification obviously cast doubt on their validity, but it is useful also to comment directly on the content and construct validity of the MTT tests.

In general, content validity refers to whether test questions cover the right material, that is some defined domain of content. For licensure and certification tests, this translates into whether test questions clearly and correctly span the domain of knowledge necessary to protect the public from people who are not competent. Since competence in most professional fields is hard to define and measure precisely, content validation studies of licensure and certification tests are usually based on the expert judgment of people in the field being tested. In the case of the MTT the relevant fields were those of education and teaching. Typically in content validation studies, test developers ask practitioners to judge whether test questions are job-related and whether they match a particular content domain (often defined in terms of test objectives).

Several Massachusetts officials have said publicly that such content validation studies have been done by panels of educators across the state in the development of the MTT. However, no relevant reports or documentation on these reviews have yet been released, even though the MTT have now been administered four times.

At the same time, a particular portion of the MTT gives us pause about the way in which the content validation and job-relatedness studies have been used in the development of the new MTT. On the first MTT, administered in April 1998, as part of the writing test, examinees were asked to transcribe a 156-word text drawn from the Federalist papers (written by James Madison in 1787) as the text was read three times by a narrator on audiotape. (According to an August 5 story in the Boston Globe, the dictation exercise was suggested by Massachusetts Board of Education members Edward Delattre and John Silber; Hart, 8/5/98).

It seems to us highly implausible that such an exercise would be judged a valid and job-related measure of writing competence by a majority of panelists reviewing content validity. Also, though we have reviewed more than 50 years of teacher competency testing in the United States (the NTE, for example was created in 1940; Haney, Madaus and Kreitzer, 1987), we have found no other instance in which a dictation exercise has been used as a measure of teacher competence.

How then could such an unusual exercise have shown up on the MTT? We cannot be sure. But it is worth noting that in the Alabama teacher testing case referred to previously (which is reproduced in part in appendix 2), Judge Myron Thompson found that on the Alabama Initial Certification Test "a significant number of items appearing on the examinations failed to reflect accurately the collective judgment of curriculum committee members. In some cases changes to actual test items were not implemented. In other cases, items that had never been reviewed by a curriculum committee appeared on examinations. . . . [Also,] many items appeared on the examinations even after they had been rated content invalid by the requisite number of Alabama panelists" (Richardson v. Lamar County Bd. of Educ. 729 F. Supp 806, 821-822). It may be recalled that the developer of the Alabama test is the same company that developed the MTT.

The construct validity of the Massachusetts Teacher Tests

A third and more general form of validity evidence of the meaning of test scores relates to the "constructs" that the scores represent. As the 1985 *Standards* point out, "Substantial relationships of a test to other measures that are purportedly of the same construct and the weaknesses of relationships to measures that are purportedly of different constructs support the identification of constructs and the differences among them" (AERA, APA & NCME, 1985, p. 10).

In carrying out test-retest analyses, we were surprised to find that in our sample of people who took the MTT in both April and July, there was a correlation of less than 0.10 between MTT reading and writing test scores in both April and July (see Table 2). Since reading and writing are both verbal or literacy skills, we would expect to find substantial correlations between test scores of these related constructs. But, as we have noted, the test-retest sample was restricted (with one odd exception) to people who had failed either the reading or writing test in April. Thus the group of repeat test- takers represents a highly restricted or attenuated sample of MTT test-takers in general.

To examine the relationship between MTT reading and writing scores on less restricted groups of test-takers, we returned to data obtained from for analyzing test-retest reliability. Several institutions had provided us with data on all of their students who had taken the MTT in April and in July. These data allowed us to examine the MTT Reading x MTT Writing correlations on larger samples than when we confined ourselves to individuals who took the MTT in both April and July. Table 7 presents results of these analyses.

Table 7: Correlations of MTT Reading and Writing Scores

	April	July
Boston College	0.42 (111)	0.20 (44)
Lesley College	0.56 (62)	0.65 (60)
Westfield State	0.57 (101)	0.50 (107)
Total N	274	211
Median r	0.56	0.50

Note: Sample sizes shown in parentheses

Note, first, that the correlations between MTT Reading and MTT Writing scores vary somewhat: from 0.42 to 0.57 for April and from 0.20 to 0.65 for July. Part of this is due surely to sample size. For example, the most anomalous correlation in Table 7 is for Boston College for July test results (correlation of 0.20). Note, however, that for this sample, there was an N of only 44. If we consider only those cases in which N>100, we see a much more consistent pattern, with MTT Reading x MTT Writing correlations of 0.42, 0.50, and 0.57. This suggests that the average correlation between MTT Reading and MTT Writing test

scores is about 0.50.

This finding may be compared with previous research on the intercorrelations between measures of two verbal skills. Cronbach (1970), for example, reports that the Verbal and Spelling subtest scores on the General Aptitude Test Battery (GATB) correlate in the range of 0.66 to 0.72. Donlon reports that Test of Standard Written English (TSWE) scores correlate with SAT Verbal scores in the range of 0.76 to 0.80 and with SAT Reading scores in the range of 0.72 to 0.77 (Donlon, 1984, p. 81). Similarly, Conrad, Trismen and Miller report that GRE Verbal and GRE Analytical scores for the same individuals correlate in the range of 0.76 to 0.77 (Conrad, Trismen & Miller, 1977, p. 19). Indeed, even SAT Verbal and SAT Mathematical scores have been found to correlate in the range of 0.64 to 0.72 (Donlon, 1984, p. 81).

These comparisons cast considerable doubt on the construct validity of the MTT Reading and Writing test scores, which correlate only in the range of 0.42 to 0.57, with an average correlation of about 0.50.

Summary

In sum, our results indicate that the MTT Reading and Writing test scores are unreliable and of doubtful validity. Specifically, we found that the scores:

- Are unreliable as indicated by our calculations of test-retest reliability (in the range of 0.50 to 0.70);
- Contain almost two to three times the degree of error as well-developed tests (with an error of measurement in the range of 9 to 17 points);
- Have high rates of misclassification (as indicated by the fact that among those who "failed" either the MTT Reading or Writing test in April, more than 50% "passed" that test in July);
- Are of questionable content validity and doubtful construct validity, as indicated by the low correlation (about 0.50) between reading and writing test scores.

Why the MTT Reading and Writing tests are so unusually unreliable and of such doubtful validity is the more mysterious because the skills of reading and writing are ones for which many reliable and valid tests have been developed over many decades. There are many possible causes for the low reliability and apparently poor validity of the MTT tests. The problems may arise from test content, administration, scoring, scaling, equating or some combination of these factors. Fortunately, another aspect of inquiry by the Ad Hoc Committee offers insight into why these scores are of such low reliability and apparently poor validity.

Interviews with MTT-Takers: Vignettes and Summary

Several individuals who sent copies of their MTT score reports to the Ad Hoc Committee spontaneously offered us comments on the new MTT. For example, one woman wrote:

> After graduation from xx College--one of the best schools in the area, ALL of my daughter's friends failed at least one section of the MA test. Is something wrong with this picture? After a phone chain among many parents, we all agree there is a problem with the MA test, as these students did all pass required testing for other states. [Underline in original; name of College deleted to protect confidentiality.]

Another correspondent wrote:

> My scores on the Praxis series earned me a license to teach Language Arts and Social studies to grades six through nine in North Carolina. Unfortunately, this was not enough to earn a reprieve from the Massachusetts test. This is just one of the aspects of the test with which I take issue.
>
> One major problem with the Massachusetts Teacher Tests is that there is no preparation offered. When I called to request information on the test, I received a packet of test objectives for each of the tests I was taking. This information was practically useless, as I still had no clue as to the format of the test. [Letter to the Ad Hoc Committee, dated August 20, 1998.]

In light of such comments, we contacted the first 15 test-takers who had sent us copies of their score reports, to ask whether they were be willing to be interviewed, on condition that we keep their identities confidential. All 15 agreed. The interviews focused on the current professional status of the teacher candidates, sought their views on the administration and content of the MTT tests and asked about their attitudes toward testing and teaching. We gathered this information in telephone interviews lasting between one half to one hour during November 1998. We took notes during phone conversations and elaborated them after the end of conversations. Typewritten accounts of each interview were then prepared, and results across interviewees were analyzed by looking for common themes and comments.

Interview sample

Although this was a small, self-selected sample, those who agreed to an interview represented a wide range of experiences. Of the fifteen, seven (47%) passed all three parts of the test (reading, writing, and subject area) on their first try, approximating the passing rate for the state overall. Two additional candidates (13 %) passed both literacy sections, but failed their subject area tests. Five (33%) passed one portion of the literacy section only, with two of them also passing their subject area. Only one out of 15 candidates interviewed failed all three portions of the test.

The 15 candidates had college degrees from nine private and four public colleges and universities, with two unknown. Although most had received a first degree in 1998, several were teachers who had moved to Massachusetts after teaching in other states, and one had 20 years experience as a teacher.

At the time of the interviews, eight of the fifteen were certified to teach in Massachusetts; eleven were certified in at least one other state--including Connecticut, New York, Rhode Island, New Hampshire, New Jersey, California, Georgia, Arkansas, Missouri, North Carolina, Maryland, and Tennessee. To receive such certification, candidates typically had passed a required test. Nine of them had taken the National Teachers Examination (NTE), recently renamed and broadened to become the Praxis. Others had taken specific state tests, the Graduate Record Examinations, or the Millers Analogies Test (MAT).

Four took the MTT in April 1998, during the test's first administration. Nine took the MTT for the first time during the second round of testing, in July 1998. Two took it in October 1998. All candidates interviewed submitted scores for the morning two-part literacy portion of the test. Candidates also submitted scores for elementary education (6), English (2), physical education (2), and general science (1), physics (1), music (1), middle school (1), and special needs (1).

Although a common rationale for teacher certification tests such as the MTT is that they will protect schools, parents and school children from incompetent teachers, the MTT tests did not prevent most teacher candidates in the sample from securing work in some kind of teaching capacity. Regardless of whether they passed or failed the MTT, 12 of the 15 candidates interviewed currently work in public, private, parochial, and charter schools, both in and out of state.

Of the seven candidates who passed the MTT, two are working in full-time teaching positions in Massachusetts public schools, while two are working in-state as long-term public school substitutes. Two more work as full-time teachers in public schools out of state. One candidate is not working by choice.

Table 8: Current Employment Status of Interviewees

	PASSED MTT (n=7)				**FAILED MTT (n=8)**		
Employment Status	Full time Teacher	Long Term Substitute	Teacher's Aide		Full time Teacher	Long Term Substitute	Teacher's Aide
Employed in education							
In Mass.							
Public School	2	2					1
Private School					1		

Charter School				1	
Out of State					
Public School		2		1	
Private or parochial school				2	
Employed, not in education			1		
Not employed	1		1		

Of the eight candidates who "failed" the MTT, six are working in schools. One works full-time in a Massachusetts charter school; a second works as a teacher in a Massachusetts private school; and a third works in the Commonwealth as a full-time public school teacher's aide. Of the remaining four, three work as full- time teachers out of state, one each in a public, private, and parochial school. One candidate is working in the travel industry, and one is not working by choice.

page 1 | introduction | background | reliability & validity | interviews | conclusions | references

Vignettes

Before summarizing the general findings from our 15 interviews, it is useful to provide vignettes of two candidates, to highlight the diversity of candidates and their experiences. Pseudonyms are used in these vignettes to maintain our agreement of confidentiality with interviewees. (Note 9)

"Peter McHugh"

Peter McHugh recently graduated summa cum laude with a 3.9 average and a major in physics from one of New England's top-ranked private colleges. His scores on the Graduate Record Exam, as documented on score reports he sent us, were 650 on the GRE- Verbal, 750 on the GRE-Quantitative and 700 on the GRE Analytical. These scores correspond to the 91st, 89th and 86th percentile among all those who took the GRE between Oct. 1, 1994, and September 30, 1997 (ETS, 1998). In contrast, he scored only 82 on the MTT Reading test. Since the DOE has issued no technical documentation on the MTT, we cannot be certain what percentile ranking would correspond with a MTT Reading test score of 82; but on the basis of MTT scores we have collected, we estimate that Peter would fall somewhere in the range of 80th to 85th percentile on the MTT Reading test--obviously quite at odds with his performance on the GRE, especially since the population taking the GRE is more selective than that taking the MTT.

Even before graduation from college, Peter was offered a job to teach at one of the country's outstanding high schools in a state that recognizes a Massachusetts teaching certificate as certifying eligibility to teach.

Although a Rhode Island resident, Peter did practice teaching in Massachusetts schools and signed up to take the Massachusetts Teacher Tests in April 1998 in the belief that the results would not "count" toward certification. Less than two weeks before the test date, when he learned he would need to pass to receive a state teaching certificate, he threw all his energy into studying for the exam. Reviewing the test guide, he found one sample question for science, but nothing for physics, and, although he was told he would receive a list of test objectives, he never received one. In the absence of any guidance to what he might encounter on the exam, Peter took the state's curriculum frameworks and studied "about 15 hours a day" for two weeks, developing his own study. Like some other candidates interviewed, Peter found testing conditions for the listening portion of the test "absolutely atrocious" and the clarity of directions "not good at all." He explained:

> On the communication and literacy sections, sometimes you would have to write in the booklet and sometimes the answer sheet. You'd go back and forth. It wasn't consistent. It was a disaster. The proctors didn't understand. It ended up being a student who figured it out and explained it to the rest of us for one section. This took about five minutes and you started the test pretty frazzled.

Peter added, "I do remember leaving the test feeling that I didn't get to show I could read or write well." In the afternoon, Peter took the physics portion of the exam. This time, his concerns were less about testing conditions and more about test content. Specifically, he found content that is not mentioned in the Massachusetts curriculum frameworks and that, in any event, he considered inappropriate for high school physics. He reported:

> There was content on semiconductors. There were graphs and charts I was supposed to analyze, and I knew nothing about semiconductors. There were two 600-1000 word essay questions at the end. One was appropriate; the second was on motors and generators. Most high school textbooks have no more than a section on this. Until the morning of the test, I had never studied motors. In most high school physics courses, you don't get to this. The questions didn't accurately reflect what is covered in a high school physics course. The content went much beyond that. The last 25% of the test had content I had never seen before. I had to skip 7 or 8 questions because they were on concepts I had never seen in 14 college courses. I filled in the bubbles because there was no penalty for guessing.

Peter is currently teaching at the out-of-state high school that hired him in May. He teaches two "Honors" and three "AP" physics classes. He passed the literacy and communications portion of the MTT, but barely passed the physics portion of the MTT. He said, "If I'd been just a little lower, who knows where I'd be now. I wouldn't have this job."

"Allegra Karnofsky"

Allegra Karnofsky, a Connecticut resident, graduated in June 1998 from a nationally-known music college based at a private four-year university out of state. Before taking the Massachusetts Teacher Test in July, Allegra had passed the Reading, Writing and Math and Music Concepts and Content portions of the Praxis, with scores required for certification in Connecticut and New Jersey.

Allegra registered for the July test knowing little more than what she had read in newspapers about the April MTT results. She reported, "What was nerve-wracking was having heard that so many people failed in April." Although the NES registration booklet noted that study guides were available, Allegra's multiple calls to the testing company and Department of Education did not produce such a guide.

Because Allegra had taken the Praxis test for music teachers, she expected the MTT would test something about her teaching ability and was surprised at how little subject matter knowledge it covered. She said:

> There was a lot missing. There was some music theory, but I felt I was taking a test of music history. I am a K-12 certified teacher in Connecticut, and nowhere do you teach music history unless it's at college. The essay questions were very broad and open-ended, and both were music history -- for example, compare a 20th century composer and an 18th century composer. But that's not music education, it's music history. Most content was on classical, and there were one or two questions on American jazz. As a music education teacher, I need to know about a lot more than classical or jazz. There was no multicultural or world music, no Latin music.

Allegra added, "It wasn't geared toward a teacher. That was what troubled me the most. The past two years have been devoted to music education, not music history."

Prior to the test, Allegra had heard about the dictation exercise on the MTT Writing test; nonetheless, she found it "terrible." She said, "The tape recording was the worst experience of my life. The tape recorder the proctor was given was poor quality. It was muffled." She added, "The dictation is strange in its own way. As a teacher, you don't have to take down what your students say."

Allegra's passing scores on the Praxis allowed her to become certified to teach in New Jersey, where she had done practice teaching, and in Connecticut, and she believes new teachers should have to pass a standardized test in order to teach. She said:

> I think all teachers should know their subject. In the Praxis test I took, you take one test, then schedule the computer-based testing when you wanted. I don't see what's wrong with the Praxis test. If Massachusetts went for the computer-based test, it would make so much more sense instead of taking a paper-pencil test.

Allegra passed the MTT Reading test with a score of 74, failed the MTT Writing test by one point, and with a score of 50 failed the Music portion of the MTT by 20 points. Although she received tickets for a retest, the tickets came with no cover letter or explanation, and she had to call the Department of Education to ask whether she had to pay again for the retest.

At the time of the interview, Allegra was teaching music at a private school in Massachusetts. Saturday classes have precluded her re- taking the Massachusetts Teacher

Test. When she called the Department of Education to ask whether the tests would be given at a time other than Saturday, she was told that alternative provisions could be made for religious reasons only. In February 1999, Allegra will begin work as a long-term substitute for kindergarten through sixth grade at a public school in Connecticut and look for a full-time job in that state. She says, "I don't know if I'm interested right now in retaking the [Massachusetts] test. I can just go on my way in Connecticut."

Summary of findings from interviews

The two vignettes above recount how two individuals experienced the MTT. These are only two of the 15 teacher candidates we interviewed. We realize fully that a sample of 15 self-selected individuals provides an extremely limited base from which to try to generalize to the experiences of all candidates who have taken the MTT. But we think it useful nevertheless to summarize some of the themes that emerged from interviews as possible causes for the low reliability and poor validity of the MTT. (A more detailed account of what was learned from the interviews appears in appendix 3.)

Many of our interviewees were dissatisfied with the information available about the MTT. Among items mentioned were the lack of a study guide, confusion over whether the April results would "count" towards certification, lack of information about conditions of retesting, and lack of detailed feedback on strengths and weaknesses of initial test performance.

A second theme among our small sample of interviewees was the conditions under which the MTT were administered. Most candidates interviewed found the general testing environment to be reasonable or on par with that of other tests taken. Nonetheless, close to half of them expressed concerns about the clarity of directions during MTT administration and about the conditions under which they performed the dictation exercise portion of the test.

Virtually all candidates interviewed mentioned the length of the MTT tests overall as excessive and believed that their 8-hour duration adversely affected their performance. Many compared the length of the MTT unfavorably with other tests they had taken and noted that the amount of writing required led to fatigue.

Teacher candidates interviewed also raised questions about the match between their "real-world" literacy skills and the test content in both the literacy and subject matter portions of the MTT. Regarding the former, candidates questioned the value of the dictation exercise and of specific questions, such as "Define a verb." As for the subject matter tests, interviewees expressed doubts about whether the tests matched Massachusetts curriculum frameworks and whether the test content was a reasonable reflection of the demands of real-world teaching. Some candidates interviewed also reported surprise that the MTT did not cover content they expected based on their experience taking other teacher tests, such as the Praxis or other states' teacher certification tests. Despite the range of concerns interviewees expressed, all agreed it was reasonable to ask teacher candidates to pass a test prior to certification. They were aware, as one put it, that "most of the professions have a test" and viewed testing one of the rites of passage into a profession. But in general, interviewees reported that the MTT compared unfavorably with other teacher certification tests they had taken.

Conclusions and Recommendations

The Ad Hoc Committee was formed in the summer of 1998 out of concern that important decisions were being based on the Massachusetts Teacher Tests (MTT) scores before any reasonable evidence had been produced concerning their reliability and validity. Since the DOE and NES have not made available any documentation on the reliability and validity of the MTT, in clear violation of professional standards concerning testing, and despite repeated requests for such documentation, the Ad Hoc Committee set out to study the technical merits of the new tests. Our original idea was compare individuals' scores on the MTT with scores on post-collegiate tests (such as the Praxis and the GRE) on which technical documentation is available. Toward this end we invited people to send us score reports on both the MTT and other tests.

As of December we had not received sufficient data to undertake a concurrent validity study, comparing MTT scores with those on established tests. But, in the meantime, we examined the reliability of the new tests. Specifically, using data on over 200 individuals who took the MTT in April and July 1998 (generously provided to us by eight institutions of higher education in the Commonwealth), we studied the test-retest reliability of the MTT. We found the correlations between April and July test to be extraordinarily low: about 0.30 for both the MTT Reading and Writing tests. Test-retest correlation coefficients for well-developed standardized tests typically range between 0.80 and 0.90. To examine the possibility that very low correlations were due to restriction of range (only people who scored below 70 on the April tests had to retake them), we corrected for attenuation due to restriction of range and estimated test-retest correlations for the unrestricted population of test-takers. The results indicated test-retest correlations of 0.50 to 0.70 -- still well below the reliability of well-developed tests.

We used these results to estimate the error of measurement in MTT scores. We found that MTT scores contain unusually high levels of measurement error--with an error of measurement on the new tests in the range of 9 to 17 points. We estimate that MTT Reading and Writing test scores contain two to three times the degree of error as well-developed tests.

Next, we compared pass and failure rates on the April and July administrations to consider the rates of misclassification on the MTT. Using both our test-retest sample, and a much larger sample of data reported on the DOE web site, we found that the MTT tests have very high rates of misclassification--as indicated by the fact that among those who "failed" either the MTT reading or writing test in April, more than 50% "passed" the test in July. Evidence suggests also that a fair number of people who "passed" the MTT did so simply because of error in the tests.

We also considered the content and construct validity of the MTT tests. At least one portion of the MTT Writing test (the dictation exercise) raises doubts about the content validity of the MTT and specifically their job-relatedness. Moreover, when we examined the correlation between MTT Reading and Writing test scores, the resulting correlations of about 0.50 raise serious doubt about their construct validity. Previous research suggests that the scores for tests of two related verbal constructs correlate in the range of 0.65 to 0.80.

Finally, we report on results of interviews with 15 candidates who took the MTT in April, July, or October (7 of whom passed and 8 failed). Since this was a small and self-selected sample, results are merely suggestive. But they indicate that the

unreliability and poor validity of MTT scores may result from the lack of a study guide for the new tests, confusion over whether the April results would "count" towards certification, poor conditions of administration (in at least some test sites), simple fatigue resulting from the 8-hour duration of the tests, and test content. Although all those interviewed supported the idea of certification testing for teachers, as is common with other professions, many compared the MTT unfavorably with other teacher certification tests they had taken (e.g. the Praxis or NTE and certification tests in other states).

Recommendations

If the Commonwealth wants high standards for its teaching force, it must use assessments that meet similarly high professional standards. The current Massachusetts Teacher Tests fail to meet this criterion. Results from the April and July administrations of the MTT reveal that the new tests are so unreliable and of such poor validity that they are passing candidates who lack the knowledge and skills the MTT are allegedly testing and failing many who do have these skills. Therefore, the Ad Hoc Committee recommends that:

1. **The Massachusetts Board of Education immediately suspend administration of the MTT.** No exam at all is better than an unreliable exam that may be mistakenly failing 50% of qualified pre- service teachers while passing unknown numbers of unqualified ones.
2. **The Commonwealth convene an independent panel of testing experts to audit the development, administration and use of the MTT in light of both of professional standards for testing and the requirements of the Education Reform Act**. These experts should issue a report evaluating how well the first four administrations of the MTT meet accepted professional standards. If they find that the MTT fails to meet these standards, they should propose other approaches that will contribute to high-quality teaching in the Commonwealth.
3. **An investigation be launched into how and why the state has allowed the new MTT tests to be used.** An independent investigation into this matter is essential, since even before contracting with NES to develop the MTT, the DOE knew that a federal court had found that same firm to have "violated the minimum requirements for professional test development" with its teacher certification tests for Alabama. That the DOE nevertheless proceeded to allow the new MTT tests to be used, in obvious violation of professional standards on testing, to make important decisions about individuals before the validity and reliability of the new tests had been documented, was a course of action so imprudent as to call out for independent scrutiny.

As James Madison wrote in 1787, in the passage candidates were asked to transcribe in the April 1998 version of the MTT, "No man is allowed to be a judge in his own cause because his interest would certainly bias his judgment and, not improbably, corrupt his integrity." So too with organizations; the DOE, having implemented new teacher certification tests of undocumented validity and reliability, should not be allowed to judge its own cause.

Notes

1. In September 1998, the Massachusetts Department of Education (DOE) announced that the name Massachusetts Teacher Tests (MTT) was being changed to Massachusetts Educator Certification Tests (MCET), to reflect the fact that not just teachers but also other professional educators, such as counselors and principals, would be required to pass the new exams. However, throughout this report we refer to the Massachusetts Teacher Tests (MTT), since that is how they are most widely known.
2. These test standards have been developed by the American Educational Research Association (AERA), the American Psychological Association (APA) and the National Council on Measurement in Education (NCME).
3. Richardson v. Lamar County Bd. of Educ. 729 F. Supp 806, M. D. Ala. 1989, p. 821. A portion of this decision appears in appendix 2 of this report.
4. NES President William Gorth signed the contract on February 23, and then Commissioner of Education Robert Antonucci on February 26, 1998.
5. The most common approaches for estimating internal consistency are the Cronbach alpha and split-half techniques.
6. We are submitting this report for publication and will make available to other investigators the complete set of data on which our reliability analyses have been based, but with the identities of the institutions of higher education removed.
7. In the remainder of this section of this report, we focus on MTT reading and writing scores. Among the more than 200 candidates whose MTT scores we obtained, there were many different subject matter tests represented. Hence the sizes of samples for any one subject matter test were much small than those for the reading and writing tests.
8. Here we should explain why we devoted considerable attention to these anomalous "outlier" scores. Such unusual cases can have a disproportionate impact on summary statistics, such as means, standard deviations and correlation coefficients. Deletion of one or two extreme cases can change the summary statistics. Hence, as we explain below, we report reliability estimates not only for our entire test-retest sample, but also for a trimmed sample from which outlier cases have been deleted. Therefore, in summary
9. Candidates whose experiences are described in these vignettes have given their consent to these descriptions. We note, however, that specific details of their cases have been altered to protect their confidentiality.

References

American Educational Research Association (AERA), the American Psychological Association (APA) and the National Council on Measurement in Education (NCME) (1985). *Standards for Educational and Psychological Testing.* Washington, DC: APA.

Anastasi, Anne (1976). *Psychological testing* (4th Edition). New York: Macmillan.

Conrad, Linda, Trismen, Donald and Miller, Ruth (Eds.) (1977). *Graduate Record Examination technical manual*. Princeton, NJ: Educational Testing Service.

Cronbach, Lee (1977). *Essentials of psychological testing* (3rd edition). NY: Harper & Row.

Donlon, Thomas (1984). *The College Board Technical Handbook for the Scholastic Aptitude Test*. New York: College Entrance Examination Board.

Educational Testing Service (1998). *GRE 1998-1999 Guide to the Use of Scores*. Princeton, NJ: Educational Testing Service.

Hambleton, Ronald K. (1999). Politicians fail, not the teachers. Education Connection (Winter issue), pp. 19-22.

Haney, W., Ludlow, L., Raczek, A., Stryker,S. and Jones, A. (1994). *Calibrating Scores on Two Tests of Adult Literacy: An Equating Study of the Test of Adult Literacy Skills (TALS) Document Test and the Comprehensive Adult Student Assessment System (CASAS) GAIN Appraisal Reading Test (Form 2)* . (Report prepared for the Manpower Demonstration Research Corporation) Chestnut Hill, MA: Boston College.

Haney, W., Madaus, G. and Kreitzer, A. (1987), Charms talismanic: Testing teachers for the improvement of American education. In E. Rothkopf (Ed.) *Review of Research in Education*. Volume 14, pp. 169-238.

Haney, W. & Raczek, A. (1993) *Surmounting outcomes accountability in education*. Washington, DC: U.S. Congress Office of Technology Assessment.

Hart, Jordana (1998). "The teachers test: Madison, via Silber." Boston Globe, 8/5/98, pp. A1 & A20

Jackson, Bailey (1998). Letter dated September 18, 1998 to Dr. David P. Driscoll, Commissioner of Education (on behalf of the Commonwealth Education Deans' Council).

Linn, Robert (1983). Pearson selection formulas: Implications for studies of predictive bias and estimates of educational effects in selected samples. *Journal of Educational Measurement*. 20:1, 1-16.

Lord, F. and Novick. M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Madaus, G. (1998). Testimony on the testing portion of Governor's Bill (H. 5677). (Testimony before the Education Committee, Massachusetts House of Representatives. July 15, 1998).

Pressley, Darrrell S. (1998). "Dumb struck: Finneran slams 'idiots' who failed teacher tests." Boston Herald, 6/26/98 pp. 1, 28.

Thorndike, Robert and Hagen, Elizabeth (1977). *Measurement and evaluation in psychology and education* (4th Edition). New York: Wiley.

Appendix 1
The Massachusetts Teacher Tests: A Chronology

- **7/85** The Massachusetts legislature passed a bill that, among other things, required candidates for teacher certification to pass a "standardized exam in his or her subject field [and] a standardized exam of communication and language skill" (MGL, 1985, Chapter 188, Section 15).

- **6/93** The Massachusetts legislature passed the Education Reform Act of 1993 which included language that required candidates to pass tests, specifically "a writing and subject matter test," in order to be certified.

- **10/96** The Board of Education had "an initial discussion about implementing ... a two-part test for teacher certification .... The Commissioner [Robert Antonucci] recommended that the Board set October 1, 1997 as the implementation date. The Board agreed to discuss this further and take action at the November meeting" (Massachusetts Department of Education, Board in Brief, 10/24/96).

- **11/96** The Massachusetts Board of Education "voted to endorse a recommendation by Commissioner Antonucci to require all candidates for teacher certification as of January 1, 1998 to pass a standardized test in communications and literacy skills and subject matter knowledge" (Massachusetts Department of Education, Board in Brief, 11/21/96). The Board also voted that Commissioner Robert Antonucci should "proceed at once with the selection of a test vendor with the aim of having the test available for review by the Board no later than October 1, 1997" (Massachusetts Department of Education, Videotape of Board Meeting of 11/18/96).

- **2/97** The DOE issued on 2/24/97 a Request for Responses (RFR) from prospective teacher certification test contractors. The RFR stated that vendors should describe in their bid 1) how they would deliver a "technical report to the Department of Education following the use of each new form of the tests with a summary for public dissemination" (p. 15); and 2) their "plan for consultation with a technical committee of nationally recognized experts recommended by the Contractor (external to the Contractor's organization) ... The technical committee will review the test items, test administration, and scoring procedures for validity and reliability and report its findings to the Department of Education" (p. 11). The RFR also asked bidding vendors to meet a timetable that included the following critical dates and events:

- **5/97-11/97** Advisory committees review sample test questions, scoring and evaluation criteria, and plan for quality control. Pilot tests conducted.

- **11/97** Test materials submitted to DOE for approval.

- **1/98-6/98** Monitoring of test quality and standardization; reliability study due.

- **12/97** Commissioner Antonucci reported to the Board on 12/18/97 that he had selected National Evaluation Systems (NES) to develop and administer the new tests, but some issues needed to be settled before executing a contract: "Issues still under discussion with the contractor include the test administration schedule and, more importantly, ensuring that the test will be rigorous and of high quality, based on college-level content" (Massachusetts Department of Education, Board in Brief, 12/15/97).

- **1/98** The DOE released a registration bulletin and an informational packet concerning these tests. The informational packet, titled "Massachusetts Teacher Tests, Questions and Answers, January, 1998" stated the following:

  > "**[Question]**: If I am now enrolled in a teacher preparation program, when should I take the tests?
  > **[Answer]**: Candidates who expect to complete their teacher preparation programs by August 31, 1998 are encouraged to take the teacher tests on either April 4 or July 11, 1998. Candidates who take the tests on these dates will satisfy the testing requirement automatically. Candidates who take the tests beginning with the October 3, 1998 administration will be required to achieve a qualifying score in order to be certified" (p. 3).

  The official 1998-1999 Registration Bulletin also informed candidates that "no qualifying score will be established until after the first two administrations of the tests ... Candidates who must take the tests and are eligible to participate in those first two administrations will satisfy the testing requirement by completing the tests" (p. 2).

- **1/98** The DOE mailed content validation surveys to school districts and teacher preparation programs asking teachers and professors to review and comment on test objectives for 31 different tests. Respondents were asked to complete and return these surveys by January 31, 1998. 4,300 eligible respondents replied.

- **2/98** Educators participated in test validation conferences for the MTT held on 2/10 and 2/12/98. Participants reviewed test bank items for: 1) match of item to test objective, 2) accuracy, 3) freedom from bias, and 4) job-relatedness (for Massachusetts teachers).

- **2/98** Robert Antonucci resigned the position of Commissioner and was later replaced by Frank Haydu as Interim Commissioner.

- **2/98** Approximately 1,500 college juniors and seniors in the state's teacher preparation programs participated in pilot testing of new open-ended and multiple-choice questions for selected tests during the early part of February.

- **2/98** The DOE and NES signed a contract to have NES develop the MTT. William Gorth signed for NES on 2/23/98; exiting Commissioner Robert Antonucci signed for the DOE on 2/26/98.

- **3/98** The DOE issued a "Study Guide" that stated the following: "This is a

preliminary edition of the Massachusetts Teacher Tests study guide. An expanded set of study guides, including sample questions and open-ended questions from each test field, will be available beginning in August 1998."

- **3/98** The DOE withdrew the study guide because the Written Mechanics Exercise, designed to assess candidates' knowledge of spelling, punctuation and capitalization, was changed. Originally, candidates were to be asked to fill in (three to five) missing portions of six different sentences that were printed in the test booklet. They were to do this as an audiotaped narrator read each sentence, several times over, in its entirety. For the April exam, however, candidates were asked to transcribe a 156-word text written by James Madison in 1787. This text, part of the Federalist Papers, was read three times by an audiotaped narrator. **(See copy of this text at the end of this chronology.)**

- **3/98** In a reversal of previous policy, the DOE announced on 3/25/98 that eligible candidates taking the April and July tests would no longer qualify automatically; instead, they would have to achieve a passing score to be provisionally certified.

- **4/98** The first round of tests were administered on 4/4/98.

- **4/98-7/98** Scoring panels met to recommend cut scores for every test.

- **6/98** The Board of Education voted on 6/22/98 to set the cut score at 1 standard error of measurement below the scores recommended by the scoring panels.

- **7/98** The Board met again on 7/1/98, at the request of Acting Governor Cellucci, and voted to raise the cut score to the level originally recommended by the scoring panels. During this meeting, Frank Haydu resigned the position of Commissioner of Education. David Driscoll was later named Acting Commissioner.

- **7/98** NES mailed score reports on 7/6/98 to April test-takers.

- **7/98** The second round of tests were administered on 7/11/98. The Ad Hoc Committee distributed flyers at five of the six test sites.

- **9/98** The DOE released a "Test Information Booklet" that contained a) sample questions from each section of the communications and literacy test; b) one sample multiple-choice question for just thirteen of the (43) subject exams; c) one sample open-response item for just one subject exam; and d) one sample oral expression and one sample written expression item from one (of the 7) foreign language exams.

- **9/98** The DOE announced that as of 9/30/98 the title of the educator certification testing program changed from Massachusetts Teacher Tests to Massachusetts Educator Certification Tests.

- **10/98** The third round of tests were administered on 10/3/98. The Ad Hoc Committee distributed flyers at all six test sites.

- **1/99** The fourth round of tests were administered on 1/9/99. The Ad Hoc Committee distributed flyers at all six test sites.

Candidates who sat for the first administration of the MTT on April 4, 1998, were asked to transcribe the following text written by James Madison in 1787. This exercise, which constituted the Written Mechanics section of the communications and literacy test, is intended to assess test-takers' knowledge of capitalization, punctuation, and spelling. An audiotaped narrator read the full text three times:

> "No man is allowed to be a judge in his own cause because his interest would certainly bias his judgment and, not improbably, corrupt his integrity. With equal, no, with greater reason, a body of men are unfit to be both judges and parties at the same time. Yet, what are many of the most important acts of legislation but so many judicial determinations, not indeed concerning the rights of single persons, but concerning the rights of large bodies of citizens? And what are the different classes of legislators but advocates and parties to the causes which they determine? It is in vain to say that enlightened statesmen will be able to adjust these clashing interests and render them all subservient to the public good. The inference to which we are brought is that the causes of faction cannot be removed, and the relief is only to be sought in the means of controlling its effects."

Appendix 2
Richardson v. Lamar County Bd. of Educ.
729 F. Supp 806 (M. D. Ala. 1989) (Excerpted)

United States District Court, M.D. Alabama, Northern Division.

Nov. 30, 1989.

MEMORANDUM OPINION

MYRON H. THOMPSON, District Judge.

Plaintiff Alice Richardson, an African-American, has brought this lawsuit claiming that defendant Lamar County Board of Education [FN1] wrongfully refused to renew her teaching contract in violation of Title VII of the Civil Rights Act of 1964, as amended. [FN2] Richardson charges the school board with two types of discrimination under Title VII. First, she asserts a claim of "disparate treatment": [FN3] that the school board refused to renew her contract because of her race. Second, she asserts a claim of "disparate impact": that the board's stated reason for not renewing her contract--that she had failed to pass the Alabama Initial Teacher Certification Test--is impermissible because the test has had a disparate impact on African-American teachers. The court's jurisdiction has been properly invoked pursuant to 42 U.S.C.A. § 2000e-5(f)(3).

FN1. Richardson has sued not only the Lamar County Board of Education but also its superintendent and members. However, because Richardson may obtain full relief from the school board the court has not treated the board members and the superintendent separately from the school board.

FN2. Title VII is codified at 42 U.S.C.A. §§ 2000e through 2000e-17.

FN3. Richardson's disparate treatment claim is also based on 42 U.S.C.A. § 1981 and the fourteenth amendment, as enforced by 42 U.S.C.A. § 1983, Jett v. Dallas Independent School District, 491 U.S. 701, 109 S.Ct. 2702, 105 L.Ed.2d 598 (1989), with jurisdiction premised on 28 U.S.C.A. §§ 1331, 1343. Because a plaintiff must prove intentional discrimination to establish a disparate treatment claim under § 1981, § 1983 and the fourteenth amendment as well as under Title VII, Stallworth v. Shuler, 777 F.2d 1431, 1433 (11th Cir.1985), and because Richardson is seeking the same relief under all these statutory provisions, the court need not address separately her theories under §§ 1981, 1983, and the fourteenth amendment. The court also need not address whether Richardson has stated a cognizable claim under § 1981. Patterson v. McLean Credit Union, 491 U.S. 164, 109 S.Ct. 2363, 105 L.Ed.2d 132 (1989).

Based on the evidence presented at a nonjury trial, the court concludes that Richardson may recover on her disparate impact claim but not on her disparate treatment claim. The court's disposition of Richardson's disparate treatment claim is simple and direct. The court simply applies the procedure set forth by the Supreme Court in Texas Department of Community Affairs v. Burdine, 450 U.S. 248, 101 S.Ct. 1089, 67 L.Ed.2d 207 (1981). The court's disposition of her disparate impact claim is, however, much more difficult.

The court first addresses and finds meritless two defenses raised by the school board: that Richardson's disparate impact claim is barred by principles of collateral estoppel and res judicata; and that under the framework set forth in Price Waterhouse v. Hopkins, 490 U.S. 228, 109 S.Ct. 1775, 104 L.Ed.2d 268 (1989), Richardson would not have been reemployed even if she had passed the state certification test. The court then goes through a lengthy application of the disparate impact analysis outlined by the Supreme Court in Wards Cove Packing Co., Inc. v. Atonio, 490 U.S. 642, 109 S.Ct. 2115, 104 L.Ed.2d 733 (1989).

I. BACKGROUND

Richardson taught in the Lamar County School System for three years, from 1983 to 1986. She was, however, unable to obtain a permanent teaching certificate and therefore had to teach with temporary and provisional certificates. To obtain a permanent certificate, Richardson, like all other teachers in the state at that time, had to *809 pass the Alabama Initial Teacher Certification Test, which consisted of a "core" examination and an examination aimed at the specific area in which the teacher sought to teach. Richardson wanted to teach in the areas of early childhood education and elementary education, and thus could meet the certification test's specific area requirement by passing the examination in either area. Between 1984 and 1986, Richardson failed the early childhood education examination twice and the elementary education examination three times.

In the spring of 1986, the Lamar County Board of Education decided that the elementary school where Richardson taught should be consolidated with another school. Because fewer teachers would be needed, the school board informed 15 nontenured teachers, including Richardson, that their contracts would not be renewed for the 1986-87 school year. Four of the 15 teachers were, however, rehired. Richardson, who would have acquired tenure if she had been rehired, was not one of the four.

Approximately a year later, in May 1987, this court enforced a consent decree requiring the State Board of Education to issue permanent teaching certificates to a court-defined class of black teachers who had failed the state teacher certification test. [FN4] Richardson received her certification pursuant to the consent decree.

FN4. Allen v. Alabama State Board of Education, 816 F.2d 575 (11th Cir.1987) (directing district court to enforce consent decree); Allen v. Alabama State Board of Education, Civil Action No. 81-697-N (M.D.Ala. May 14, 1987) (enforcing the consent decree).

[NOTE: Omitted from this reproduction of Judge Thompson’s opinion are several pages in which he discussed: II. DISPARATE TREATMENT CLAIM; and III. DISPARATE IMPACT CLAIM. The remainder of the opinion is reproduced in its entirety.]

[15] Since Richardson has established that the early childhood education and elementary education examinations had an adverse racial impact, the burden shifts to the Lamar County Board of Education to produce evidence of employment justification. An understanding of the history of the Alabama Initial Teacher Certification Test is important to determining whether the school board has met its burden and, if so, whether Richardson has, in turn, shown that the school board's justification for the certification test has no basis in fact.

a. History of the Early Childhood Education and Elementary Education

Examinations

In 1979, amidst a national groundswell in favor of teacher competency testing, the Alabama State Board of Education placed development of a uniform certification test at the head of its agenda. It retained a professor at Auburn University to conduct a feasibility study regarding implementation of a teacher testing program in Alabama; the state's Assistant Superintendent for Teacher Certification also participated in the study. After a rather cursory investigation, the two educators recommended implementation of a testing program similar to one designed by a private test developer for the State of Georgia.

The State Board agreed with the recommendation. In January 1980, it awarded a contract to the private test developer on a noncompetitive basis. [FN29] While the board did not always express its purpose for imposing the test requirement with perfect clarity, both the test developer and the board understood that the test would measure whether a teacher possessed enough minimum content knowledge to be competent to teach in the classrooms of Alabama.

FN29. Board members anticipated that the test requirement would adversely impact against African-American applicants for teaching certificates. However, the same decision would have been reached without consideration of that factor. The board's action was predicated on a legitimate concern for improving the quality of education in Alabama.

The time frame for development of the Alabama Initial Teacher Certification Test, as it came to be known, was quite short. The test developer had one year to complete development and implementation of 36 separate examinations. The test developer created a "core" examination and 35 additional examinations that covered specific subject areas. As stated, a teacher had to pass the core examination and one subject area examination in order to receive certification.

The Assistant State Superintendent, the sole ranking state official charged with oversight of the private test developer's contract compliance, had a doctorate in educational administration; but neither he nor anyone on his staff had any expertise in test development. And no outside experts were retained to monitor the test developer's work. The developer's work product was accepted by the state largely on the basis of faith.

The test developer began by preparing a preliminary planning document. It next asked the State Department of Education to appoint Alabama educators to the various committees and panels necessary for completion of the project. According to criteria provided by the developer, these educators were selected to represent a fair cross section of persons from different geographic areas throughout the state. They were also selected in such a way that African- Americans and women were fairly represented overall; however, not all committees and panels had minority representatives.

The test developer's technical staff and subject area consultants then formulated topic outlines for the various examinations. They consulted state education standards, state courses of study, materials related to Alabama's student competency tests, and examples

of textbooks used in Alabama public schools. They also developed actual test objectives. These objectives were more explicit statements of concepts embedded in the topic outlines. The objectives were reviewed by the developer's editors and management. The developer's in-house work was far below average.

*818 In October of 1980, approximately 200 Alabama educators attended a two-day conference to review the topic outlines and objectives for 36 examinations. They had previously been mailed orientation materials. After additional orientation, they were divided into curriculum committees to review the topic outlines for comprehensiveness, organization, accuracy, and absence of bias. The committees then reviewed the objectives to ensure that they matched the topic outlines. Taxonomic level, significance of content, accuracy, level of specificity, suitability, and lack of bias were considered. Decisions were reached by consensus during both stages of review. Modifications and deletions were recorded by the test developer's personnel assigned to each committee. In some cases, however, the developer made additional changes, or ignored suggested changes, without obtaining clearance from committee members. No effort was made at any time to link the topic outlines and objectives to the state-mandated curriculum for teacher training programs.

The test developer then sent a job analysis survey packet to approximately 3,000 in-service teachers throughout Alabama. The purpose of this survey was to determine the job relatedness of the test objectives. [FN30] However, in nine fields where there were fewer than 200 teachers throughout the state, the test developer's process resulted in very small response rates. The survey packet was sent to persons certified and teaching in specific content areas. The packet included a set of objectives for that content area, a survey form, and a set of instructions. The teachers were asked whether they had taught or used each objective in the past two school years. If the answer was yes, they were asked to rate the objective in terms of time and essentiality. The scales used to record those responses were balanced in favor of indicating that an objective was job related, and teachers were instructed to resolve doubts in favor of job relatedness. The results of the job analysis survey were tallied in such a way that responses from only those who indicated that an objective had been used in the last two years were reflected in the data. Those who indicated that an objective had not been used were ignored.

FN30. A stratified random sampling technique was employed to select survey respondents and a fair cross section of teachers was generally achieved.

In January of 1981, the curriculum committees met for a second time. They were provided results from the job analysis survey and were asked to determine which objectives should generate questions to appear on the examinations. This step was called "objective selection." The survey results were a major determinant of which objectives were ultimately selected.

The test developer then prepared a "blueprint" for each examination. These blueprints specified the number of test questions, or items, necessary to measure each objective. Test items were drafted by the test developer's content area consultants and edited by its staff. Again, the developer's in- house work was far below average.

In March of 1981, the test items were reviewed by Alabama curriculum committees for "item/objective" match, significance of content, accuracy, clarity, and absence of bias. This "item review" process lasted for two days. Committee revisions were recorded by

the test developer's personnel. However, in some cases, the developer ignored the suggested changes, or made additional recisions, without consulting committee members for approval. In other cases, the developer simply added new items that had never been reviewed by committee members. As many as 20 items for each 120-question examination fell into one of these categories.

In late April of 1981, the test developer convened a separate group of educators to review the test items once again for content validity. The purpose of this session, which lasted one day, was to provide an independent check against the judgments already rendered by the previous committees of Alabama educators. The new panelists reflected a fair cross section of persons in their field and were qualified to make content validity judgments in their *819 field. Each educator worked separately, but votes were tallied as if educators had served on a committee. After orientation, the educators were asked to judge whether each item matched its objective, was accurate, was free of bias, and was not tricky, misleading, or ambiguous. If the item met these criteria, the item was rated content valid by that judge. If the item was deemed invalid, the judge's reason for rejecting that item was recorded. The test developer compiled these content validity ratings; a level of agreement among judges greater than 50% was required for an item to be deemed content valid. While a majority of items appearing on the final test instruments reflected the judgment of Alabama educators that those items were content valid, a significant number of items appearing on the tests did not reflect that judgment. These included those items that had been revised by the developer without obtaining clearance from the panelists. [FN31]

FN31. The test developer did not convene separate panels of minority educators at any stage of the item review or content validity process to screen items for possible bias.

The judges were also asked to make cut-score decisions for those items they had rated content valid. For these items, and those items only, judges were asked whether a teacher with minimum content knowledge in the field should be able to answer the item correctly. A yes-no response was requested. Judges were disqualified from making that same cut-score determination for any item they had previously rated content invalid. In essence, their expert judgment as to those items was ignored.

The test developer then assembled and produced the actual test instruments for all 36 examinations. Each examination had 100 items tentatively designated as scoreable and 20 items tentatively designated as nonscoreable. The examinations were first administered to a group of actual candidates. The test developer had originally contemplated a separate field tryout, but time constraints prohibited such a course. After the first administration, the developer examined item statistics to flag problem questions. Based on this item analysis, it selected 100 scoreable items and 20 nonscoreable items for each examination. The developer did not conduct empirical bias studies to determine whether the difficulty of items varied according to the race of examinees.

The test developer then set a minimum cut score for each examination. The developer's original plan was to take the panelists' cut-score ratings and subject them to a 10% non-cumulative binomial algorithm. This level of agreement among judges would then determine the minimum cut score. However, the developer's procedure yielded cut scores that were so astoundingly high that they signaled, on their face, an absence of correlation to minimum competence. For example, of the more than 500 teachers who took the first administration of the core examination, none would have passed if the

original cut-score methodology had been followed.

Faced with this problem, the test developer made various mathematical "adjustments" to the original cut score. First, the developer applied a 10% cumulative binomial algorithm. When the cut scores still remained too high, it applied a 5% cumulative binomial algorithm. This process of applying successively stricter algorithms was referred to at trial as a "binomial twist." The developer engaged in this process without consulting the State Department of Education or any Alabama educators. In two fields--that of Music and that of Speech, Communication, and Theatre--the 5% binomial twist yielded cut scores that were much too low. The developer simply applied a different mathematical algorithm to those examinations; again, the developer consulted no one. For all special education and school counseling examinations, the developer recommended a uniform cut score cap of 80 to the State Department of Education. This recommendation was based on the developer's experience in the Georgia testing program. However, in Georgia, the decision to place a cap on cut scores was reached by state officials in conjunction with Georgia educators. [*820 FN32]

FN32. Although the cut scores in the special education area were intended to serve as an upper limit, the cut scores on five of those examinations were actually raised to 80 to achieve uniformity.

The State Department of Education was then given the option of dropping the cut scores, as set by the developer, by two or three standard errors of measurement (SEM's). It was clear at that time that cut scores, even after the various adjustments catalogued above, were not measuring competence. For example, even after the developer's 5% binomial twist, 78% of the teachers taking the first administration of the core examinations would have failed. The same would have been true for 93% of those taking the school counseling examination, 89% of those taking the learning disability examination, and 97% of those taking the library media examination. Instead of challenging what the developer had done, the state simply dropped the cut scores three SEM's in order to arrive at a "politically" acceptable pass rate. In so doing, the state knew that the examinations were not measuring competency.

In 1982, the test developer formulated nine additional examinations. Its test construction procedures and quality of execution were essentially the same, with the following exceptions. First, the developer's job analysis survey form contained a rating scale with additional errors. Second, a more restrictive binomial table was used to calculate agreement among panelists on content validity questions. Third, a more accurate cut-score methodology was employed.

In 1983, the developer conducted a "topicality review" to update ten of the examinations already in use. A curriculum committee performed item and objective review. The committee's tasks were to determine whether items had become stale because of changes in the teaching field and to identify problems with items by reference to item statistics for the first eight administrations of the certification test. On average, 50% of the items in any given examination were replaced or revised. The developer did not convene a separate panel, as it had during the initial test development, to provide an independent screen for content validity, nor was an independent cut-score panel convened. The curriculum committee provided ratings used to set cut scores.

b. Validity of the Early Childhood Education and Elementary Education

Examinations

The Lamar County Board of Education contends that the state teacher certification test was designed to determine whether a teacher is competent to teach in Alabama's classrooms. Richardson claims, as stated, that the early childhood education and elementary education examinations were invalid, that they did not measure competency.

Generally, validity is defined as the degree to which a certain inference from a test is appropriate and meaningful. APA Standards at 94. [FN33] It is suggested that validity evidence must necessarily be restricted to success on the job; and, to be sure, there are Title VII decisions that have approached the question of validity by asking whether a given score on a test yields an appropriate and meaningful inference about successful performance on the job. See, e.g., Contreras v. City of Los Angeles, 656 F.2d 1267, 1271-1272 (9th Cir.1981), cert. denied, 455 U.S. 1021, 102 S.Ct. 1719, 72 L.Ed.2d 140 (1982); Guardians Association of New York City Police Dept., Inc. v. Civil Service Commission, 630 F.2d 79, 91 (2d Cir.1980), cert. denied, 452 U.S. 940, 101 S.Ct. 3083, 69 L.Ed.2d 954 (1981). However, there is no magic to using success on the job as an anchor point for validity. Success on the job is just one of many constructs that a test can measure. Thus, a sound inference as to a different construct, such as minimal competence, may also form the basis for a finding of validity. In short, a test will be valid so *821 long as it is built to yield its intended inference and the design and execution of the test are within the bounds of professional standards accepted by the testing industry. APA Standards at 9; cf. Washington v. Davis, 426 U.S. 229, 247 & n. 13, 96 S.Ct. 2040, 2051 & n. 13, 48 L.Ed.2d 597 (1976) (validity need not be limited to inference about success on the job).

FN33. The term APA Standards is a shorthand reference for the American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Standards for Educational and Psychological Testing (1985).

In order to be valid, a licensure or certification test must support the inference that persons passing the test possess knowledge necessary to protect the public from incompetents. APA Standards at 63. Part of an appropriate validation strategy for licensure and certification tests is to define clearly and correctly the domain of minimum content knowledge necessary for competence. The test domain, once defined, must then be translated into actual test questions that measure competence. At all stages, validity flows from the expert judgment of practitioners in the field being tested. The test developer's role is to employ professionally accepted practices that accurately marshal the expert judgment of those practitioners. When the questions on a given test actually measure what practitioners in the field consider to be content knowledge associated with competency, the test instrument is held to possess content validity. However, mere content validity does not alone establish test validity. No matter how valid the test instrument, an inference as to competence or incompetence will be meaningless if the cut score, or decision point, of the test does not also reflect what practitioners in the field deem to be a minimally competent level of performance on that test. Again, the test developer's role in setting a cut score is to apply professionally accepted techniques that accurately marshal the judgment of practitioners.

In assessing the overall validity of the Alabama Initial Teacher Certification Test, the court must therefore address both content and cut-score validity. The test developer

retained by the State Board of Education followed a multi- step procedure to build 36 teacher certification examinations in 1981. With minor variations, it followed the same procedure when it built nine additional examinations in 1982. The developer then applied a third procedure when it revised ten examinations in 1983. The content validity of each of these examinations turns on whether the developer's procedures were adequate, or were outside the bounds of professional judgment. For reasons that follow, the court concludes that the developer's procedures violated the minimum requirements for professional test development. Accordingly, none of the examinations, including the early childhood education and elementary education examinations, possesses content validity.

The test development process was outside the realm of professionalism due to the cumulative effect of several serious errors committed by the developer when it formulated the 45 examinations in 1981 and 1982. First, while practicing teachers were asked to offer their judgment about the job relatedness of test objectives, it is clear that the test developer's survey instrument distorted that judgment. Scales were balanced in favor of finding job relatedness and respondents were specifically instructed to resolve all doubts in favor of job relatedness. Moreover, the response of those teachers who indicated that they had not used an objective was ignored.

Second, Alabama educators serving on curriculum committees selected test objectives based on those survey results. It has been suggested that the survey was used only in an advisory capacity and that any survey errors were offset by the overall judgmental process undertaken by committee members. However, it is plain that the survey was conducted to solicit critical firsthand knowledge from in-service teachers. It is equally plain that curriculum committee members, aware that the survey had been conducted for that purpose, took the survey results quite seriously. The court concludes that the overall judgmental process for determining job relatedness of test objectives was distorted significantly by survey error.

Third, a significant number of items appearing on the examinations failed to reflect accurately the collective judgment of curriculum committee members. In some *822 cases, changes to actual test items were not implemented. In other cases, items that had never been reviewed by a curriculum committee appeared on examinations. It is suggested that, in any testing program of this size, a certain number of errors of this type will be found. The court agrees with this proposition in principle; however, the evidence reflects that the error rate per examination was simply too high.

Fourth, Alabama educators were never asked to determine whether the test items themselves were job related, even though such an approach is standard practice in the testing industry.

Fifth, many items appeared on the examinations even after they had been rated content invalid by the requisite number of Alabama panelists. It is suggested that, before any such item appeared on a final test form, it was revised by the test developer, and that all revisions were approved by Alabama panelists. However, neither the State Board of Education nor the test developer produced any documentation of this alleged revision and approval process. Moreover, not a single panelist was called at trial to confirm that the process had actually occurred. The court finds that no such process occurred and that the test developer simply substituted its own judgment for that of Alabama educators.

In 1983, the test developer conducted a topicality review for ten of the examinations already in use. It is suggested that, even if those ten examinations were previously content invalid, they gained content validity by way of the topicality review process. The court does not agree. The topicality review process resulted in changes to, or replacement of, only about 50% of any given examination's 120 items. Items that were not revised or replaced therefore remained just as invalid as they were at birth. Moreover, as to items that were revised or replaced, there was no separate content validity determination. The court agrees with Richardson's experts that, on balance, these two factors rendered the ten examinations subjected to the 1983 topicality review to be content invalid as well. [FN34]

FN34. The court does not agree that the test developer's multi-step test development process was inherently self-correcting. There is substantial support in the record for the view that errors at one step not only survived the next step, but also created new errors.

Moreover, the fact that a validity study for the National Teachers Examination was upheld in United States v. South Carolina, 445 F.Supp. 1094 (D.S.C.1977), aff'd 434 U.S. 1026, 98 S.Ct. 756, 54 L.Ed.2d 775 (1978), does not mandate the same result here. The validity of the present examinations must be assessed on the basis of evidence now before the court. Cf. York v. Alabama State Board of Education, 581 F.Supp. 779, 786 (M.D.Ala.1986) ("tests are not valid or invalid per se ...; the fact that the validity of a particular test has been ruled upon in prior litigation is not necessarily determinative in a different factual setting").

Richardson advances an array of challenges to the cut-score methodology employed by the test developer. It is clear that, as to the 35 examinations developed in 1981, the cut scores bear no rational relationship to competence as that construct was defined by Alabama educators. [FN35] The *823 evidence reveals a cut-score methodology so riddled with errors, that it can only be characterized as capricious and arbitrary. There was no well- conceived, systematic process for establishing cut scores; nor can the test developer's decisions be characterized as the good faith exercise of professional judgment. The 1981 cut scores fall far outside the bounds of professional judgment.

FN35. The court must point out that three of Richardson's arguments with respect to the 1981 cut scores clearly lack merit. First, she asserts that Nassiff's 1978 "Two-Choice Angoff" method for yielding an original cut score was and is "without professional endorsement." However, professional literature published well before the initiation of Alabama's testing program endorsed methodologies similar to Nassiff's approach. See R. Thorndike, Educational Measurement at 514-515 (1971). Moreover, while current professional literature does not grant Nassiff's method the highest possible marks, it certainly does not condemn it as being wholly outside the bounds of professional judgment. See Berk, A Consumer's Guide to Setting Performance Standards on Criterion Referenced Tests, 56 Rev. of Educ. Research 137, 148 (1986). Second, Richardson contends that Nassiff's method was largely unproven and that an alternative cut-score methodology should have been used at the same time as a backup. While the court agrees that this might have been advisable, there is no evidence that the failure to use a backup cut-score method was unprofessional. Third, Richardson argues that the test developer's recent adoption of a more sophisticated cut-score methodology signals the bankruptcy of Nassiff's 1981 method. The court does not agree. The fact that, with new developments in the field, the test developer later changed its methodology should not

be held against it as an admission of error.

First and foremost, it is undeniable that cut scores for the 35 examinations developed in 1981 do not reflect the judgment of Alabama educators who served as panelists on the minimum cut score committees. This is a crucial error, because competence to teach is a construct that can only be given meaning by the judgment of experts in the teaching profession. Here, expert panelists who rated an item invalid as to content were automatically disqualified from going on to indicate whether that item should be counted toward the minimum cut score. This means that when a panelist indicated that an item should be excluded--because it contained inaccurate content, did not measure an objective, was tricky, ambiguous, or misleading, or was biased--that panelist's opinion was ignored for purposes of determining whether the item measured competence and should contribute to the cut score. The exclusion of such opinions resulted in a series of cut scores that reflected a distorted notion of competence.

Second, the court has no doubt that, after the results from the first administration of those 35 examinations were tallied, the test developer knew that its cut-score procedures had failed. The proof of this fact is that none of the more than 500 teachers who took the first administration of the core examination would have passed if the original cut score, calculated according to the developer's original plan, had been utilized. The court cannot conclude that all Alabama teachers who took that examination were totally and completely incompetent. It follows, therefore, that the developer knew that its cut-score procedure had utterly failed to reflect a valid construct of competence.

Third, instead of notifying the State Department that its cut-score procedure had malfunctioned, the test developer attempted to mask the presence of system failure by making various unilateral mathematical "adjustments" to the original cut score until an "acceptable" score had been reached. The most common adjustment was application of a "binomial twist" to the data collected from Alabama educators. This adjustment tended to lower cut scores. It is argued that lowering cut scores offset any system failure that might have occurred previously. This argument, however, misses the mark. The critical factor with respect to cut-score validity is not whether there was a net change in cut-score level, but whether the cut score itself accurately reflected the expert judgment of Alabama educators about whether examinees possess the competence to teach. This construct of competence cannot be guessed at by out-of-state test makers. It is also argued that the developer's resort to the "binomial twist" was an exercise of "tempered judgment" in light of actual examination data. Again, however, the fatal error is that it was the developer, and not Alabama educators, that exercised this judgment. [FN36]

FN36. It is argued that the binomial twist was, in fact, implemented in consultation with the State Department of Education, and that such consultation somehow injects the judgment of Alabama educators into the cut-score process. However, the evidence is clear that the developer never consulted any official at the State Department of Education with respect to the binomial twist. In fact, the State Department was not advised of that twist until shortly before trial.

Fourth, in two fields--that of music, and and that of speech, communication and theatre--the 5% binomial twist yielded cut scores that were much too low. In those areas, the developer simply applied a different mathematical algorithm to arrive at an acceptable cut score. Again, the developer substituted its judgment about competence for that of Alabama educators.

Fifth, for all special education and school counseling examinations, a uniform cut score of 80 was adopted. To be sure, the *824 State Department of Education made this decision, based on a policy judgment that no score should exceed 80. However, it is clear that the developer played an advisory role in that decision and that its advice was completely irresponsible. The developer recommended holding the scores at 80 based on its experience in the Georgia testing program. However, in Georgia, the decision to place a cap on cut scores was reached by the State Department of Education in conjunction with Georgia educators. The test developer never suggested that the State Department consult Alabama educators, and there is no evidence that such consultations in fact occurred. In effect, the developer assumed that the judgment of Georgia educators in a different testing program would be good enough for the people of Alabama. Once again, cut scores bore no relation to the expert judgment of Alabama educators. Moreover, if the rationale for adopting a cut score of 80 was to place a cap on such scores, it is difficult to understand why the cut scores for five special education examinations were actually raised to 80.

Sixth, the State Board did not drop the cut scores, as set by the developer, to advance bona fide psychometric or policy purposes. The board did not drop the scores three SEM's to account for measurement error; the developer recommended a drop of only two SEM's for that purpose. Nor were scores dropped three SEM's to reduce adverse impact against blacks; the State Assistant Superintendent in charge of the certification test was vehemently opposed to taking race into account in setting the cut scores. Finally, while cut scores may have been lowered by three SEM's in part for the permissible purpose of maintaining an adequate teacher supply, the court is convinced that the primary purpose for dropping three SEM's was to mask the obvious system failure generated by the developer's cut-score methodology. For example, even after the developer's binomial twist, 78% of the teachers taking the first administration of core examinations would have failed, and the same would have been true for 93% of those taking the school counseling examination, 89% of those taking the learning disability examination, and 97% of those taking the library media examination. It is apparent that these pass rates did not reflect a fair construct of minimal competence. Further adjustments were employed to back into a passing rate that would appear tolerable and reasonable. The State Board of Education and the test developer in effect abandoned their cut-score methodology, with the result that arbitrariness, and not competence, became the touchstone for standard setting.

The court would be inclined to uphold the cut-score procedures employed for the nine examinations developed in 1982 and the ten examinations subjected to topicality review in 1983; however, each of these examinations has already been shown to be content invalid. Since a valid cut score cannot be generated by items that lack content validity, the validity of the cut-score procedure itself is not enough. Accordingly, the cut scores for the 1982 and 1983 examinations are also invalid.

In reaching the above conclusions, the court has been sensitive to a number of factors. First of all, as stated earlier, close scrutiny of any testing program of this magnitude will inevitably reveal numerous errors, and these errors will not be of equal footing. Secondly, cut scores cannot be determined with mathematical certainty, and political considerations may properly enter into cut-score decisions. The court's task therefore is to assess the sum gravity of the defects found, and to determine whether, as a result of these defects, the examinations are invalid as to content and cut scores. The court

recognizes that, in carrying out this task, it must proceed with caution, and even deference. Although the court must assess the credibility of testimony advanced by each side and arrive at an independent judgment, the court should not readily set aside the findings of those who developed a test; the mere fact that the court sees things differently should not, by itself, be considered sufficient to impeach such findings. But while a court should eschew an idealistic view of test validity, it should also be careful not to apply an "anything *825 goes" view. In other words, the mere presence of conflict in expert testimony does not prove that a test fails to meet minimum standards; nor does it prove that a test meets such standards. A court should find a test invalid only if the evidence reflects that the test falls so far below acceptable and reasonable minimum standards that the test could not be reasonably understood to do what it purports to do. The court is convinced that this was the case with the Alabama Initial Teacher Certification test, and in particular with the early childhood education and elementary education examinations. [FN37]

FN37. The court recognizes that it has focussed not so much on the early childhood education and elementary education examinations, but on the Alabama Initial Teacher Certification Test as a whole. The court has done this because the history of the two examinations challenged by Richardson is the same as the history of the teacher certification test as a whole; the conclusions reached by the court regarding the certification test are also applicable to the two challenged examinations. Moreover, in order to appreciate fully the invalidity of the two challenged examinations, one must also understand just how bankrupt the overall methodology used by the State Board and the test developer was.

The court also recognizes that it has focussed on the development and implementation of several individual examinations which have not been challenged by Richardson. The court has included these examinations as additional evidence of the invalidity of the State Board and test developer's overall methodology.

IV. RELIEF

Since Richardson is entitled to prevail on her disparate impact claim, the court must now determine her relief. The court will require that the Lamar County Board of Education reemploy Richardson as an elementary school teacher at a salary and with such employment benefits and job security as would normally accompany the position had she been employed in the school system since 1983. The court will also require that the school board pay her all backpay and other employment benefits she would have received had the school board reemployed her for the 1986-87 school year. The court will also require that the school board pay reasonable attorney's fees to her attorney. 42 U.S.C.A. § 2000e-5(k). The court will give Richardson and the school board an opportunity to agree, between themselves, to the appropriate amount of attorney's fees, present pay, backpay, and other employment benefits to which Richardson is entitled. If the parties are unable to agree, the court will then set these matters down for a hearing.

An appropriate judgment will be entered.

JUDGMENT AND INJUNCTION

In accordance with the memorandum opinion entered this date, it is the ORDER, JUDGMENT, and DECREE of the court:

1. That judgment be and it is hereby entered in favor of plaintiff Alice Richardson and against defendants Lamar County Board of Education and its superintendent and members;
2. That it be and it is hereby DECLARED that plaintiff Richardson may recover on her "disparate impact" claim but not on her "disparate treatment" claim against defendants Lamar County Board of Education and its superintendent and members;
3. That defendants Lamar County Board of Education and its superintendent and members, their officers, agents, servants, employees, attorneys, and those persons in active concert or participation with them who receive actual notice of this injunction by personal service or otherwise, be and they are each hereby ENJOINED and RESTRAINED from failing to reemploy forthwith plaintiff Richardson as an elementary school teacher in the Lamar County School System at a salary and with such employment benefits and job security as would normally accompany the position had she been employed in the school system since 1983;
4. That plaintiff Richardson be and she is hereby awarded from defendants Lamar County Board of Education and its superintendent and members all backpay and other employment benefits she would have received had said defendants not illegally refused to reemploy her;
5. That plaintiff Richardson and defendants Lamar County Board of Education *826 and its superintendent and members be and they are hereby allowed 21 days from the date of this order to file a request for the court to determine the appropriate amount of present pay, backpay and other employment benefits to which plaintiff Richardson is entitled, should the parties be unable to agree to these matters;
6. That plaintiff Richardson be and she is hereby allowed 28 days from the date of this order to file a request for reasonable attorney's fees; and
7. That all other relief sought by plaintiff Richardson that is not specifically granted be and it is hereby denied.

It is further ORDERED that this court retains jurisdiction of this case until further order.

It is further ORDERED that all costs of these proceedings be and they are hereby taxed against defendants Lamar County Board of Education and its superintendent and members, for which execution may issue.

The clerk of the court is DIRECTED to issue a writ of injunction.

Appendix 3
Summary of Results of Interviews with Examinees

Our interviews with fifteen candidates who had sent us copies of their MTT score reports yielded background on the information available to test takers, administration of the test, including testing conditions, test content, test length, and test-takers attitudes about testing requirements for professional certification. In part 4 of this report, we summarized the nature of candidate views on these topics. Here we provide more detail on what was said by how many of our sample of 15.

Information available to test takers

According to teacher candidates interviewed, neither the testing company, National Evaluation Systems (NES), nor the Massachusetts Department of Education, provided adequate information about the test prior to administration. Regardless of whether they took the MTT in April, July or October, not one of the 15 teacher candidates interviewed reported receiving any useful information (or information they considered useful) from the Department of Education. Nine (60%) saw no study guide at any time. Two July test-takers called the Department of Education repeatedly and still failed to receive any information. Others reported receiving information that was misleading or inadequate, either regarding the consequences of the test or about test content or format.

Candidates reported that when they called the Massachusetts Department of Education, they were told no study guide was available. Of the candidates interviewed, all who contacted the DOE found that contact unsatisfactory. They described this contact and the information provided on the test as "not particularly helpful," "not readable," "vague" and "very limited."

Information for first-time test-takers

April test-takers registering for the test believed that test results would not "count" toward certification in Massachusetts. They reported that their colleges and universities also believed that because this was the first administration of the test, their scores would be used only to determine passing scores for future test-takers. This understanding was based on more than hearsay. The NES "1998-1999 Registration Bulletin" that many first-time candidates received reported:

> No qualifying score will be established until after the first two administrations of the tests on April 4 and July 11, 1998. Candidates who must take the tests and are eligible to participate in those first two administrations will satisfy the testing requirement by completing the tests. A qualifying score for each test will be determined by fall 1998 and used beginning with the October 3, 1998 administrations. From then on, candidates for provisional or provisional with advanced standing teacher certification who must take the tests will have to achieve a qualifying score to meet the certification requirement (p. 2).

One candidate prepared her own study guide based on Massachusetts curriculum

frameworks, noted, "The six of us who graduated were not told much. Less than two weeks before the test date we were told we needed to receive a certain score to get a certificate."

July test-takers reported they received most of their information about the test from the media. Because of widespread media reports, some reported they were able to anticipate certain test components, specifically the dictation portion used to test "literacy and communication skills." In September 1998, the State Department of Education released a guide entitled "Massachusetts Teacher Tests: Test Information Booklet." The booklet includes test objectives, sample questions, and criteria for test scoring. This resource was unavailable to April and July test takers.

In the view of some candidates interviewed, the confusion about the test's consequences contributed to depressing test results. Whether they passed or failed, spring test- takers in particular attributed low test scores overall to the message that the test would not "count." One April test-taker noted, "They had told us the scores weren't going to matter. I thought it was a joke." Another reported:

> Many of us were in the middle of student teaching. We were really busy, and most people didn't study so much as I did. This could explain why so many people failed the first time around and then passed the next time.

A third candidate who passed all three portions of the test nonetheless remarked:

> I would guess that a number of people who took the very first test took it in panic mode. Because of all the hype, a lot of people went into the test in a panicked frame of mind. Under these circumstances, test scores are not an accurate representation of what they can do.

Information about Re-testing

Candidates interviewed also reported receiving inadequate information, if any, regarding a re-test for those who failed the test on the first round. For example, test- takers knew that free re-tests were offered to all candidates who failed one or more sections of the April test. However, one July test-taker reported that she did not know if such provisions would be made for those who failed the July test. Another candidate who failed the April test but did not intend to teach in Massachusetts described her uncertainty as to whether a re-test was still required because she had graduated from a private college in Massachusetts. She reported, "I was on the phone all the time. No one was giving me a definite answer."

Information about test results

Several candidates observed that they received individual test results in a format that precluded them from learning from their mistakes. One who passed the test on the first round in April asserted, "It is not surprising at all that there are repeated re-tests. There is no information at all about what you did wrong, no information about what you did right." In part, these test-takers' wish for more precise information about their performance reflected some doubt about the objectivity of the questions and their scoring. As one noted:

> I'd like to see what I got wrong. A lot of questions were subjective and poorly worded. I could have supported any of my answers, even if they were marked incorrect.

Given the lack of either a study guide or individual feedback, some candidates were left with the belief that their results were a matter of chance. For example, one experienced teacher who passed still found the morning literacy portion of the test "very subjective." As she explained:

> For example, [on a question requiring the] summarizing of an article. I thought I did a good job, but my "bar" said I did an "adequate to inadequate" job on this. You could have three people grade this and come up with three different scores.

A July test-taker added:

> [There were] a couple of open-ended questions on how you would teach something. But different people could teach it in different ways. There was more than one right answer. There was no guidance, no rubrics. I would have liked to have seen a study guide.

Some candidates who passed the MTT on the second round attributed their scores as much to luck as to preparation. These candidates saw little difference between the knowledge they possessed at the time of their first test, compared with that knowledge at the time of their second test. Pressed to explain the differences, they alluded to differences in testing conditions, including more time to take specific sections of the test, familiarity with test format and content, or luck. For example, questioned as to what they thought had contributed to raising test scores from one administration to the next, one candidate who had failed the first round on the elementary portion of the test by two points replied, "It's a matter of luck." Another responded, "I don't know. Chance, totally chance!"

Testing Conditions

Candidates interviewed had taken the Massachusetts Teacher Test at seven different locations around Massachusetts. These testing sites included Bunker Hill Community College, Auburn High School, Burncote High School in Worcester, Malden High School, West Springfield High School, Randolph High School, and Wakefield Memorial High Schools. We asked the candidates to describe the testing conditions they experienced as they related to physical comfort, breaks allowed, adequacy of time to complete the test, acoustics, clarity of directions, and variations available for candidates with disabilities.

Most candidates generally found the general testing environment to be reasonable or on par with other tests taken. Problems at specific locations were noted by four candidates, including two July test-takers who mentioned heat and lack of air-conditioning as a problem, one who described proctors who talked "very loudly" among themselves for 20 minutes during the testing, and one who reported the test started late. However, almost half (47%) reported serious problems with acoustics that hindered performance on the written mechanics portion of the test. In addition, almost half also described problems

related to the clarity of directions and test organization. Others raised concerns about the length of the test overall.

Acoustics

The morning portion of the MTT purports to test written mechanics by asking candidates to listen to a passage on a tape recorder, then write down the passage word for word. The purpose of this exercise is to "demonstrate the ability to spell, capitalize, and punctuate according to standards of edited American English" (Massachusetts Teacher Tests: Test Information Booklet, 1998, p. 9). Despite the apparently straightforward nature of the task, candidates reported a number of problems in its administration. Although eight candidates tested at four of the sites described the testing acoustics as "fine" or "very quiet," other candidates described conditions that hindered test performance. Five candidates (33.0%) tested at three different locations reported hearing tape recordings from other testing rooms. One test-taker voiced a common complaint:

> The acoustics were absolutely atrocious. The classrooms were adjacent to each other. I was near the back. I couldn't hear the tape recorder in my classroom, but I could hear the tape in the classroom behind me, and they were not in sync.

Clarity of test directions and format

In addition to variations in and problems associated with the quality of listening conditions, some candidates described problems with the clarity of test directions and test format. Specifically, while six candidates characterized the proctors' understanding and test's clarity of directions as "fine," "okay," or "not bad," nine others described proctors as "poorly prepared" or "kind of lost," and directions as "vague."

Several April test-takers noted that the lack of consistency in the test's format was a problem. As one explained, "You went from multiple choice to essay to fill- in-the-blank to correct a sentence that was incorrect." A July test-taker described confusing directions in relation to test items on the elementary education sub-test. For example, she explained that when she came to one test item, she was "not sure if they wanted a lesson plan or an essay." By the third administration, test organization and clarity of directions still raised concerns. As one October test-taker reported:

> I had a problem with [the clarity of directions]. They were extremely disorganized.... Because there were so many subtests...you had the blue section of Test A ...it was confusing, with too many subtests, colors, page numbers. It was terribly confusing. I don't see why they couldn't have had one test booklet. Also there was a misprint. They had misprinted the direction of where to answer something. They said it was on one page, but it was on another.

Two test-takers also described confusion related to the "tricky" selection of answers for multiple choice items. For example, one candidate estimated that up to 70% of the special education questions offered choices like "1 and 3 only," "2 and 1 only," "3 only," and "4 only."

Length of test

Whether they had passed or failed the first time they were tested, virtually all candidates mentioned the length of the test overall as excessive and believed it had an impact on their performance. One noted, "It was too long being in one room." Another reported, "Time was the big issue. Eight hours was just too long."

In particular, several candidates mentioned how the test's emphasis on writing essay questions or copying down dictated material contributed to increasing fatigue or incoherence over the course of the day. One candidate, an experienced teacher certified to teach high school English in New York, New Hampshire, and Connecticut and who described himself as a "good test-taker," compared the MTT to other state tests and reported:

> In other tests, a third was definitely related to education, a third was about literature, and a third had to do with analytical or problem solving skills. The New York test had only one essay. In Massachusetts, there was one in the morning, and two in the afternoon, plus the morning dictation. There was a lot of writing. I was used to it, but my handwriting's awful. By the time you get to the end of the day, you're not legible.

Another candidate, an experienced elementary teacher with teaching certification in two other states added:

> There was a lot of writing in terms of essays...summary of a story, essays on different topics, two more essays on elementary ed.... It was such a long process. Having the writing at the end was hard. In the end, I was exhausted. By the time I got to the end of the content test in elementary ed, I was just filling in the bubbles. My knowledge of the elementary subject area is much higher than I scored.

Second-time test takers faced some relief from the demands of an eight-hour testing day. Those taking only one or even two portions of the MTT as a re-test consistently reported that "it was easier" or "it was shorter," leaving them less tired. Candidates interviewed who took a re-test reported benefiting from the opportunity to re-rest under less pressure and with more time allowed. For example, one candidate who scored 60 on the writing portion of the test in April, passed with a score "in the high 80s" in July. Asked how she would account for her improved score, she reported:

> I knew I would be out in two hours because I was taking only the writing. I think the only thing I can think of is that I wasn't about to pass out. Maybe my handwriting was better because I

only wrote for two hours.

Another April-test taker who passed the writing portion of the test on the second try reported no difference either in her approach to the test or in the format or difficulty of the questions. However, she noted, "It was a lot shorter. I only had to take one part. Mine was the 8:00 to 10:00 slot, so I was done for the day, and I was not exhausted." Some repeat test-takers also noted that on the re-test, they had the entire morning to complete a section that was allotted only two hours the first time around.

In commenting on the impact of the overall length of the MTT, candidates compared the MTT unfavorably to the length of other post-graduate tests. For example, one candidate who had also taken Graduate Record Examinations in her subject area noted that she had taken the GRE in four and half hours. Others noted that because it was possible to take the National Teachers Examination in sections, it was not so tiring. As one elementary teacher explained:

> I took the pre-professional skills tests for the NTE on three separate days. You had different portions, but usually you'd come in for testing at 8:00 and be done by 10:30. NTE was offered every weekend, so I could break it up into different days.

Literacy and Communication Skills Content

Teacher candidates had a number of questions about the match between their "real world" literacy skills and test content. Whether they passed or failed the test, candidates reported that they found some of the content perplexing. Without any study guide or information from others who had taken the test before them, the expectations of April test-takers in particular were shaped primarily by other post-graduate tests. Given this, first-time test- takers were especially surprised at the dictation portion of the test, and several who had taken other teacher tests noted they had never encountered anything like this on any other examination. Generally, candidates were confused about what skills and knowledge the dictation portion of the test was attempting to assess. One successful October test-taker commented:

> The dictation section was mindless, a complete waste of time. They could have just given us the section and had us punctuate it. You don't have to write it down word for word. A monkey can do that.... It was very simple-minded.

Others also questioned other portions related to reading and writing. As one successful candidate reported:

> The reading and writing content didn't seem appropriate. [It was] not testing my knowledge of reading or writing. Like the question, "Define a verb." The test didn't seem to be testing what it said it was testing.

Subject Area Content

Candidates also detailed concerns about the content in the subject area tests. Their comments detailed concerns that much of the content did not match content found in the Massachusetts frameworks or the demands of real-world teaching responsibilities.

Content not included in grade-level curriculum frameworks

Three of the candidates, all of whom passed the test on the first round, believed the content tested extended beyond the knowledge they needed for teaching. One candidate mentioned that content included went well beyond what she would expect to teach her middle grades students. She asserted, "There should be separate tests for middle and high school science. Never will I teach advanced physics in middle school."

Another who took the English subject area test questioned the extent to which some content tested even fit within the boundaries of the discipline. As he commented:

> I distinctly remember a test question on "new journalism." I see this as a separate field. The question had to do with who was responsible for the rise of new journalism. I have a BA and an MA, and sometimes journalism students came into our classes, but we didn't take theirs. I've studied criticism, but not journalism.

Lack of content about professional knowledge

Some candidates interviewed reported surprise that the MTT did not cover content they expected based on their experience taking other teacher tests, whether the National Teachers Examination (Praxis) or state tests. Most often, candidates mentioned the lack of content related to professional knowledge and skills and commented on the lack of attention to teaching itself. For example, one recent graduate and physical education teacher noted:

> The content wasn't fair. There was not a lot of application to how you taught. The test did not have, "In this situation, what would you do?" kinds of questions.

An experienced elementary teacher certified in two fields in another state likewise wondered how well the MTT could determine teachers' competence in classrooms given the mismatch she perceived between the demands of teaching and content on the MTT. She asserted:

> The MTT did not test my ability to teach. As a teacher, I'm constantly reteaching myself. There are so many things you do as teacher that are performance- based, not knowledge-based. You need to know teachers are performing, handling their class in a professional manner.

Another teacher with twenty years experience and certification in two other states, also questioned what she perceived as an overall stress on mechanical skills and content knowledge compared to the limited emphasis on knowledge of classroom practice. She reported:

> There was a big lack of anything to do with classroom management. It was *all* content knowledge. There were no more than a couple of questions on

> teaching methods, classroom management, anything to do with your ability to teach.

In light of limited content geared toward assessing professional knowledge, candidates raised concerns that the emphasis on content to the exclusion of other aspects of teaching reflected a misunderstanding of what was required to be a successful teacher. As one candidate with an exemplary academic record asserted:

> Nothing on the whole test addressed the issue of how well you teach. Even if I'd gotten every question right, it wouldn't have proved I could teach physics to 16-year-olds.

From the perspective of another candidate, even when test items ostensibly melded content knowledge with teaching knowledge, the questions did not provide adequate information to answer them well. For example, one experienced teacher described one question on the middle school portion of the test as unrelated to real-world teaching conditions:

> You were supposed to create a unit plan with a team, but in the test, you're on your own. Then you're told you're not being graded on the creativity or usability of your plan. But there are no references, resources, or curriculum frameworks to work with. You're told your response will be graded on your knowledge of your subject area, not on the lesson plan.

Candidates viewed the weighting of basic literacy and writing skills over teaching knowledge, regardless of the subject area, as a major flaw in the test. As one explained:

> It's possible with the Massachusetts test for some very good teachers to be knocked out because of problems with spelling. A lot of what we do [in classrooms] is done ahead of time. It's not a handicapping situation not to know everything there is to know. It's also possible that a lot of bad teachers are passing. People could pass the content test without knowing how to teach.

Views of teacher testing

Despite the range of concerns raised about the MTT, the candidates interviewed did not object to standardized testing per se, and all agreed it was reasonable to ask teacher candidates to pass a test prior to certification. However, they added that although they were not opposed to the testing of new teachers, they believed that the MTT should be replaced with a different test. Whether they passed or failed, candidates interviewed were aware that "most of the professions have a test" and viewed testing one of the rites of passage into a profession. However, candidates doubted that the Massachusetts test in particular could adequately assess teacher competence. They compared the MTT unfavorably to other professional tests they had taken and described the latter as having a more balanced focus on all aspects of teaching, including content knowledge, classroom skills, and problem solving. As one noted, "Yes, I'm in favor of a fair test like those offered in Connecticut or New York. But this one is not fair at all."

Candidates noted in particular that, compared with other tests, items pertaining to the teaching process itself were missing from the Massachusetts test. As one explained:

> In New York, you have to pass a basic liberal arts and sciences test.... But in New York you have to pass a written assessment of teaching skills. This includes subject knowledge but also more teaching. There are also two essays associated with classroom situations.

A third elaborated:

> It's not unreasonable to have new teachers take a standardized test. It's only unreasonable to take an unproven test that has no validation data. This one is much too subjective. The New York test was more fair in assessing teachers' ability to teach.... If they want to know if you can write under pressure, that's what this test shows.

Eight candidates explicitly mentioned the National Teacher Examination as a more accurate test of teaching skills. One experienced teacher who taught elementary school in two other states before moving to Massachusetts asserted, "The Massachusetts Teacher Test needs to look into the NTE. It's recognized, accepted, reliable." Another described the NTE as testing "a wider range of content-- communications, professional knowledge, and general knowledge." Another concluded, "Massachusetts could do better, could have a much better quality test."