February 2001

Educational policy analysis archives

Arizona State University

University of South Florida

Education Policy Analysis Archives

Volume 9 Number 6 **February 22, 2001** **ISSN 1068-2341**

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

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*.

Teacher Test Accountability: From Alabama to Massachusetts

Larry H. Ludlow
Boston College

Abstract

Given the high stakes of teacher testing, there is no doubt that every teacher test should meet the industry guidelines set forth in the *Standards for Educational and Psychological Testing*. Unfortunately, however, there is no public or private business or governmental agency that serves to certify or in any other formal way declare that any teacher test does, in fact, meet the psychometric recommendations stipulated in the *Standards.* Consequently, there are no legislated penalties for faulty products (tests) nor are there opportunities for test takers simply to raise questions about a test and to have their questions taken seriously by an impartial panel. The purpose of this article is to highlight some of the psychometric results reported by National Evaluation Systems (NES) in their *1999 Massachusetts Educator Certification Test (MECT) Technical Report*, and more specifically, to identify those technical characteristics of the MECT that are inconsistent with the *Standards*. A second purpose of this article is to call for the establishment of a standing test auditing organization with investigation and sanctioning power. The significance

of the present analysis is twofold: a) psychometric results for the MECT are similar in nature to psychometric results presented as evidence of test development flaws in an Alabama class-action lawsuit dealing with teacher certification (an NES-designed testing system); and b) there was no impartial enforcement agency to whom complaints about the Alabama tests could be brought, other than the court, nor is there any such agency to whom complaints about the Massachusetts tests can be brought. I begin by reviewing NES's role in *Allen v. Alabama State Board of Education*, 81-697-N. Next I explain the purpose and interpretation of standard item analysis procedures and statistics. Finally, I present results taken directly from the *1999 MECT Technical Report* and compare them to procedures, results, and consequences of procedures followed by NES in Alabama.

Teacher Test Accountability: From Alabama to Massachusetts

From its inception and continuing through present administrations, the Massachusetts Educator Certification Test (MECT) has attracted considerable public attention both regional and around the world (Cochran-Smith & Dudley- Marling, in press). This attention is due in part to two disturbing facts: 1) educators seeking certification in Massachusetts have generally performed poorly on the test, and 2) in many instances politicians have used these test results to assert, among other things, that candidates who failed are “idiots” (Pressley, 1998).

The purpose of the MECT is “to ensure that each certified educator has the knowledge and some of the skills essential to teach in Massachusetts public schools” (National Evaluation Systems, 1999, p. 22). The Massachusetts Board of Education has raised the stakes on the MECT by enacting plans to sanction institutions of higher education (IHEs) with less than an 80% pass rate for their teacher candidates (Massachusetts Department of Education, 2000). One consequence of this proposal is that most IHEs are considering requirements that the MECT be passed before students are admitted to their teacher education programs. In addition, Title II (Section 207) of the Higher Education Act of 1998 requires the compilation of state “report cards” for teacher education programs, which must include performance on certification examinations (U.S. Department of Education, 2000).

What all of this means is that poor performance on the MECT could prevent federal funding for professional development programs, limit federal financial aid to students, allow some IHEs be labeled publicly “low performing”, and prove damaging at the state-level when states are inevitably compared to one another upon release of the Title II report cards in October 2001. Given the personal, institutional, and national ramifications of the test results, there is no question that the MECT should be expected to meet the industry benchmarks for good test development practice as set forth in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999). At this time, however, there is no public or private business or governmental agency either within the Commonwealth of Massachusetts or nationally that can certify or in any other formal way declare that the MECT does (or does not), in fact, meet the psychometric recommendations stipulated in the *Standards.* The National Board on Educational Testing and Public Policy (NBETPP) serves as an “independent organization that monitors testing in the US” but even it does not function as a regulatory agency

(NBETPP, 2000).

In addition to the absence of a national regulatory agency, many state departments of education do not have the professionally trained staff to answer directly technical psychometric questions. Nor do they usually have the expertise on staff to confront a testing company, which they have contracted, and demand a sufficient response to a technical question raised by outside psychometricians. Furthermore, even when a database with the candidates' item- level responses is available for internal analysis, a state department of education does not typically conduct rigorous disconfirming analyses, e.g. evidence of adverse impact. Thus, most state departments are largely dependent on whatever information testing companies decide to release. The public is then left with an inadequate accountability process.

One purpose of this article is to highlight some of the psychometric results reported by National Evaluation Systems in their *1999 MECT Technical Report* (NES, 1999). Specifically, this article identifies technical characteristics of the MECT that are inconsistent with the *Standards*. A second purpose of this article is to voice one more call for the establishment of a standing test auditing organization with powers to investigate and sanction (National Commission on Testing and Public Policy, 1990; Haney, Madaus & Lyons, 1993).

The significance of the present analysis is twofold. First, psychometric results reported by NES for the MECT are similar in nature to psychometric results entered as evidence of test development flaws in an Alabama class- action lawsuit dealing with teacher certification (*Allen v. Alabama State Board of Education*, 81-697-N). That suit was brought by several African-American teachers who charged, among other things, that “the State of Alabama's teacher certification tests impermissibly discriminate[d] against black persons seeking teacher certification;” the tests “[were] culturally biased;” and the tests “[had] no relationship to job performance” (*Allen*, 1985, p. 1048). Second, there was no impartial enforcement agency to whom complaints about the Alabama tests could be brought, other than the court, nor is there any such agency to whom complaints about the Massachusetts tests can be brought. These two points are linked in an interesting and troubling way--NES, the Massachusetts Educator Certification Tests contractor, was also the contractor for the Alabama Initial Teacher Certification Testing Program (AITCTP).

Some of the criticism of debates about teacher testing, teacher standards, teacher quality, and accountability suggests that arguments are, in part, ideologically, rather than empirically based (Cochran-Smith, in press). This may or may not be the case. This article, however, takes the stance that regardless of one's political ideology or philosophy about testing, the MECT is technically flawed. Furthermore, because of the lack of an enforceable accountability process, the public is powerless in its efforts to question the quality or challenge the use of this state-administered set of teacher certification examinations. In this article I argue that the consequences of high-stakes teacher certification examinations are too great to leave questions about technical quality solely in the hands of state agency personnel, who are often ill- prepared and under-resourced, or in the hands of test contractors, who may face obvious conflicts-of-interest in any aggressive analyses of their own tests.

In the sections that follow, I begin by reviewing NES's role in *Allen v Alabama*. Then I explain the purpose and interpretation of standard item analysis procedures and statistics. Finally I compare results taken directly from the *1999 MECT Technical Report* with statistical results entered as evidence of test development flaws in *Allen v Alabama*.

NES and the AITCTP

Allen, et al. v. Alabama State Board of Education, et al.

In January 1980, National Evaluation Systems was awarded a contract on a non-competitive basis for the development of the Alabama Initial Teacher Certification testing Program (AITCTP). Item writing for these tests began in the Spring of 1981, and the first administration of the tests took place on June 6, 1981. *Allen v Alabama* was brought just six months later on December 15th, 1981. The Allen complaint challenged the Alabama State Board of Education's requirement that applicants for state teacher certification pass certain standardized tests administered under the AITCTP. On October 14, 1983, class certification (Note 1) was granted, and the first trial was set for April 22, 1985. Subsequent to a pre-trial hearing on December 19, 1984 and "after substantial discovery was done,"(Note 2) an out-of-court settlement was reached on April 4, 1985. A Consent Decree was presented to the U.S. District Court April 8, 1985(Note 3). The Attorney General for the State of Alabama immediately "publicly attacked the settlement" (*Allen*, 1985, p. 1050), claiming that it was illegal. Nonetheless, the consent decree was accepted by the court October 25, 1985 (*Allen*, Oct. 25. 1985). A succession of challenges and appeals on the legality and enforceable status of the settlement resulted (Note 4). For example, on February 5, 1986, the district court vacated its October 25th order approving the consent decree (*Allen*, February 5, 1985, p. 76). While the plaintiffs appeal of the February 5th decision was pending at the 11th Circuit Court of Appeals, trial began in district court on May 5, 1986.

The AITCTP consisted of an English language proficiency examination, a basic professional studies examination, and 45 content-area examinations. The purpose of the examinations was to measure "specific competencies which are considered necessary to successfully teach in the Alabama schools" (*Allen*, Defendants' Pre-Trial Memorandum, 1986, p. 21). A pool of 120 items for each exam was generated--100 of which were scorable and mostly remained unchanged across the first eight administrations. Extensive revisions were incorporated into most of the tests at the ninth administration. By the start of the May 1986 trial the tests had been administered 15 times in all.

A team of technical experts (Note 5) for the plaintiffs was hired in November 1983 (prior to the ninth administration of the exams) to examine test development, administration, and implementation procedures. The team was initially unsure about the form of the sophisticated statistical analyses they assumed would have to be conducted to test for the presence of "bias" and "discrimination", the bases of the case. That is, the methodology for investigating what was then called "bias" and is now called "differential item functioning" was far from well established at that time (Baldus & Cole, 1980). Nevertheless, when the plaintiffs' team received the student-level item response data from the defendants, their first steps were to perform an "item analysis." Such an analysis produces various item statistics and test reliability estimates. These initial analyses produced negative point-biserial correlations. Although point-biserial correlations are explained in detail below, suffice it to say at this point that it was a surprise to find negative point-biserial correlations between the responses that examinees provided on individual items and their total test scores. Such correlations are not an intended outcome from a well-designed testing program.

These statistical results prompted a detailed inspection of the content, format, and answers for all the individual items on the AITCTP tests. Content analyses yielded discrepancies in the keyed correct responses in the NES test documents and the keyed correct responses in the NES- supplied machine scorable answer keys (i.e., miskeyed

items were on the answer keys). This finding led to an inspection of the original NES in-house analyses which revealed that negative point-biserials for scorable items existed in their own records from the beginning of the testing program and continuing throughout the eighth administration without correction.

What this meant for the plaintiffs was that NES had item analysis results in their own possession which indicated that there were mis-keyed items. Nonetheless they implemented no significant changes in the exams until they were faced with a lawsuit and plaintiffs' hiring of the testing experts to do their own analyses. The defendants argued that it was normal for some problems to go undetected or uncorrected in a large-scale testing program because the overall effect is trivial for the final outcome. The problem with that argument was that many candidates were denied credit for test items on which they should have received credit, and some of those candidates failed the exam by only one point. In fact, as the plaintiffs argued, as many as 355 candidates over eight administrations of the basic professional skills exam alone should have passed but were denied that opportunity simply because of faulty items that remained on the tests (Milman, 1986, p. 285). It should be noted here that these were items that even one of the state's expert witnesses for the defense admitted were faulty (Millman, 1986, p. 280).

Establishing that there were flawed items with negative point-biserial correlations was critical to the plaintiffs' case. The plaintiffs presented as evidence page after page of so-called “failure tables” (Note 6) with the names of candidates for each test whose answers were mis-scored on these faulty items. Based upon these failure tables, any argument from defendants that the mis-keyed items did not change the career expectations for some candidates would most likely have failed.

In the face of this evidence, the defendants argued at trial that

> ...the real disagreement is between two different testing philosophies. One of these philosophies would require virtual perfection under its proponents' rigid definition of that word. The other looks at testing as a constantly-developing art in which professional judgment ultimately determines what is appropriate in a particular case”
> (*Allen*, Defendant's Pre-trial Memorandum, 1986, p. 121-2).

Plaintiffs counter-argued

> “This case…is not a philosophical case at all. This case is a case on professional competence….this was an incompetent job, unprofessional, and as I said before, sloppy and shoddy, and in the case of the miskeyed items, unethical.” (Madaus, 1986, p. 185).

Judge Thompson, in the subsequent *Richardson* decision which also involved the AITCTP, specifically agreed with plaintiffs on this point (*Richardson*, 1989, p. 821, 823, 825). Excellent reviews of the diametrically opposed plaintiff and defendant positions may be found in Walden & Deaton (1988) and Madaus (1990).

At the same time that this case was proceeding, the plaintiffs' appeal to reverse the vacating of the original settlement was granted prior to a decision in this trial (*Allen*, Feb. 5, 1986, p. 75). The U.S. Court of Appeals decided the district court should have enforced the consent decree (*Allen*, April 22, 1987)—which the district court so ordered on May 14, 1987 (*Allen*, May 14, 1987). Although the decision to uphold the original settlement was a positive ruling for the plaintiffs, it also was somewhat counter-productive for them because it was unexpectedly beneficial to NES at this stage in the proceedings. That is because the evidence presented above in *Allen v Alabama*

was critical of the state and NES (NES was explicitly referred to in the court documents). Thus, NES's best hope for avoiding a written opinion critical of their test development procedures was if plaintiffs' appeal were to be upheld and the original settlement enforced, as it was. Then there would be no evidentiary record, no court ruling, and no legal opinion that would reflect badly upon the NES procedures. *Richardson v Lamar County Board of Education* (87-T-568-N) commenced, however, and the actions of NES and the Alabama State Board of Education were openly discussed and critiqued in the court's opinion of November 30, 1989 (though NES was not mentioned by name in the *Richardson*, 1989 decision).

Richardson v Lamar County Board of Education, et al.

Like *Allen v Alabama*, *Richardson v Lamar County* also addressed issues of the "racially disparate impact" of the AITCTP (*Richardson*, 1989, p. 808). The Honorable Myron H. Thompson again presided, and testimony from *Allen v Alabama* was admitted as evidence (*Richardson*, 1989). Although the defendants denied in the *Allen v Alabama* consent decree that the AITCTP tests were psychometrically invalid, and even though no decision was reached in the abbreviated *Allen v Alabama* trial, the State Board of Education did not attempt to defend the validity of the tests in *Richardson v Lamar* and, "in fact, it conceded at trial that plaintiff need not relitigate the issue of test validity" (*Richardson v Alabama State Board of Education*, 1991, p. 1240, 1246).

Judge Thompson's position on the test development process of NES was clearly stated: "In order to fully appreciate the invalidity of the two challenged examinations, one must understand just how bankrupt the overall methodology used by the State Board and the test developer was" (*Richardson*, 1989, p. 825, n. 37). While sensitive to the fact that "close scrutiny of any testing program of this magnitude will inevitably reveal numerous errors," the court concluded that these errors were not "of equal footing" and "the error rate per examination was simply too high" (*Richardson*, 1989, pp. 822- 24) Thus, none of the examinations that comprised the certification test possessed content validity because of five major errors by the test developer and the test developer had made six major errors in establishing cut scores (*Richardson*, 1989, pp. 821-25).

Case Outcomes in Alabama

The *Allen v Alabama* consent decree required Alabama to pay $500,000 in liquidated damages and issue permanent teaching certificates to a large portion of the plaintiff class (*Allen*, Consent Decree, Oct. 25, 1985, pp. 9-11). The decree also provided for a new teacher certification process. However, no new test was developed or implemented and the Alabama State Board of Education suspended the teacher certification testing program on July 12, 1988. In 1995 the Alabama State Legislature enacted a law requiring that teacher candidates pass an examination as a condition for graduation. Subsequently, another trial was held February 23, 1996 to decide the state's motions to modify or vacate the 1985 consent decree (*Allen*, 1997, p. 1414). Those motions were denied on September 8, 1997 (*Allen*, Sept. 8, 1997). Given the rigorous test development and monitoring conditions of the Amended Consent Decree, it was estimated by the court that the State of Alabama would not gain complete control of its teacher testing program "until the year 2015" (*Allen*, Jan. 5, 2000, p. 23). Only recently has a testing company stepped forward with a proposal for a new Alabama teacher certification test (Rawls, 2000).

Plaintiff Richardson was awarded re-employment, backpay, and various other

employment benefits (*Richardson*, 1989, pp. 825-26). Defendants (the State of Alabama and its agencies) in both cases were ordered to pay court costs and attorney fees (*Richardson*, 1989, pp. 825-26). However, even though NES was responsible for the development of the tests, NES was not named as one of the defendants in these cases and was not held liable for any damages (Note 7).

Psychometric and Statistical Background

At this point it is appropriate to discuss some of the psychometric concepts and statistics that are fundamental to any question about test quality. The purpose of this discussion is to illustrate that excruciatingly complex analyses are not necessarily required in order to reveal flaws in a test or individual test items. The first steps in test development simply involve common sense practice combined with sound statistical interpretations. If those first steps are flawed, then no complex psychometric analysis will provide a remedy for the mistakes.

One of the simplest statistics reported in the reliability analysis of a test like the MECT is the “item-test point-biserial correlation.” This statistic goes by other names such as the “item-total correlation” and the “item discrimination index.” It is called the point-biserial correlation specifically because it represents the relationship between a truly dichotomous variable (i.e., an item scored as either right or wrong) and a continuous variable (i.e., the total test score for a person). A total test score, here, is the simple sum of the number of correctly answered items on a test.

The biserial correlation has a long history of statistical use (Pearson, 1909). One of its earliest measurement uses was as an item-level index of validity (Thorndike, et al., 1929, p. 129). The “point”-biserial correlation appeared specifically for individual dichotomous items in an item analysis because of concerns over the assumptions implicit in the more general biserial-correlation (Richardson & Stalnaker, 1933). It was again used as a validity index. It subsequently came to acquire diagnostic value and was re-labeled as a discrimination index (Guilford, 1936, p. 426).

The purpose of this statistic is to determine the extent to which an individual item contributes useful information to a total test score. Useful information may be defined as the extent to which variation in the total test scores has spread examinees across a continuum of low scoring persons to high scoring persons. In the present situation, this refers to the extent to which well qualified candidates can be distinguished from less capable candidates.

Generally, the greater the variation in the test scores, the greater the magnitude of a reliability estimate. Reliability may be defined many ways through the body of definitions and assumptions known as Classical Test Theory or CTT (Lord & Novick, 1968). According to CTT, an examinee's observed score (X) is assumed to consist of two independent components, a true score component (T) and an error component (E). One relevant definition of reliability may be expressed as the ratio of true-score variance to observed- score variance. Thus, the closer the ratio is to 1.0, the greater the proportion of observed-score variance that is attributed to true-score variance.

The KR-20 reliability estimate is often reported for achievement tests (Kuder & Richardson, 1937, Eq. 20, p. 158). Although reliability as defined above is necessarily positive, the KR-20 can be negative under certain extraordinary conditions (Dressel, 1940) but typically ranges from 0 to +1. Nevertheless, the higher the value, the more “internally consistent” the items on a test. The magnitude of the KR-20, however, is

affected by the direction and magnitude of the point-biserial correlations. Specifically, total test score reliability is decreased by the inclusion of items with near-zero point-biserial correlations and is worsened further by the inclusion of items with negative point-biserial correlations. This is because each additional faulty item increases the error variance in the scores at a faster rate than the increase in true-score variance.

Technically, the point-biserial correlation represents the magnitude and direction of the relationship between the set of incorrect (scored as "0") and correct (scored as "1") responses to an individual item and the set of total test scores for a given group of examinees. In other words, it is a variation of the common Pearson product-moment correlation (Lord & Novick, 1968, p. 341). It can range in magnitude from zero to . An estimate near zero is a poorly discriminating item that contributes no useful information. An estimate of +1 would indicate a perfectly discriminating item in the sense that no other items are necessary on the test for differentiating between high scoring and low scoring persons. A value of 1.0 is never attained in practice nor is it sought (Loevinger, 1954). Negative estimates are addressed below.

Ideally the test item point-biserial correlation should be moderately positive. Although various authors differ on what precisely constitutes "moderately positive", a long-standing general rule of thumb among experts is that a correlation of .20 is the minimum to be considered satisfactory (Nunnally, 1967, p. 242; Donlon, 1984, p. 48) (Note 8). There is, however, no disagreement among psychometricians on the direction of the relationship—it has to be positive.

The direction of the correlation is critical. A positive correlation means that examinees who got an item right also tended to score above the mean total test score and those who got the item wrong tended to score below the mean total test score. This is intuitively reasonable and is an intended psychometric outcome. Such an item is accepted as a good "discriminator" because it differentiates between high and low scoring examinees. This is one of the fundamental objectives of classical test theory, the theory underlying the development and use of the MECT.

A negative point-biserial correlation, however, occurs when examinees who got an item correct tended to score below the mean total test score while those who got the item wrong tended to score above the mean total test score. This situation is contrary to all standard test practice and is not an intended psychometric outcome (Angoff, 1971, p. 27). A negative point-biserial correlation for an item can occur because of a variety of problems (Crocker & Algina , 1986). These include:

1. chance response patterns due to a very small sample of people having been tested,
2. no correct answers to an item,
3. multiple correct answers to an item,
4. the item was written in such a way that "high ability" persons read more into the item than was intended and thus chose an unintended distracter while the "low ability" people were not distracted by a subtlety in the item and answered it as intended,
5. the item had nothing to do with the topic being tested, or
6. the item was mis-keyed, that is, a wrong answer was mistakenly keyed as the correct one on the scoring key.

When an item yields a negative point-biserial correlation, the test developer is obligated to remove the item from the test so that it does not enter into the total test score calculations. In fact, the typical commercial testing situation is one where the test contractor administers the test in at least one field trial, discovers problematic items,

either fixes the problems or discards the items entirely, and then readministers the test prior to making the test fully operational. The presence of a flawed item on a high-stakes examination can never be defended psychometrically.

One additional point must be made. The point-biserial correlation can be computed two ways. The first way is to correlate the set of 0/1 (incorrect/correct) responses with the total scores as described above. In this way of computing the statistic, the item for which the correlation is being computed contributes variance to the total score, hence, the correlation is necessarily magnified. That is, the statistical estimate of the extent to which an item is internally consistent with the other items "tends to be inflated" (Guilford, 1954, p.439).

The second way in which the correlation may be computed is to compute it between the 0/1 responses on an item and the total scores for everyone but with the responses to that particular item removed from the total score (Henrysson, 1963). This is called the "corrected point-biserial correlation." It is a more accurate estimate of the extent to which an individual item is correlated to all the other items. It is easily calculated and reported by most statistical software packages used to perform reliability analyses (e.g., SPSS's Reliability procedure).

Various concerns have been raised over the interpretation of the point-biserial correlation because the magnitude of the coefficient is affected by the difficulty of the item. The fact is, however, that all the various discrimination indices are highly positively correlated (Nunnally, 1936; Crocker & Algina, 1986). Furthermore, even though the magnitude of the point-biserial correlation tends to be less than the biserial-correlation, all writers agree on the interpretation of negative discriminations. "No test item, regardless of its intended purpose, is useful if it yields a negative discrimination index"(Ebel & Frisbie, 1991, p. 237). Such an item "lowers test reliability and, no doubt, validity as well" (Hopkins, 1998, p. 261). Furthermore, "on subsequent versions of the test, these items [with negative point-biserial correlations] should be revised or eliminated (Hopkins, 1998, p. 259).

NES AND THE MECT

The 1999 MECT Technical Report

In July 1999 NES released their five volume *Technical Report* on the Massachusetts Educator Certification Tests. Volume I describes the test design, item development description, and psychometric results. Volume II describes the subject matter knowledge and test objectives. Volume III consists of "correlation matrices by test field." Volume IV consists of various content validation materials and reports. Volume V consists of pilot material, bias review material, and qualifying score material. The report was immediately hailed by Massachusetts Commissioner of Education David P. Driscoll: "I have said all along that I stand by the reliability and validity of the tests, and this report supports it." (Massachusetts Department of Education, 1999).

Field Trial

Technical Report Volume I contains the psychometric results for the first four administrations of the MECT (April, July, and October 1998, and January 1999). It does not, however, contain any results from a full-scale field trial, nor are any "pilot" test results reported (Note 9). There is no information on how may different items were tested, where the items came from, how many items were revised or rejected, what the

revisions were to any revised items, or what the psychometric item-level results were. In fact, there is no field trial evidence in support of the initial inclusion of any of the individual items on the operational exams because *there was no field trial*.

Interestingly, the Department of Education released a brochure in January 1998 stating that the first two test administrations would not count for certification—implying that the tests would serve as a field trial. Chairman of the Board of Education John Silber, however, declared in March 1998 that the public had been misinformed and that the first two tests would indeed count for certification. This policy reversal was unfortunate because of the confusion and anxiety it created among the first group of examinees and because it prevented the gathering of statistical results that could have improved the quality of the test.

NES had considered a field trial of their teacher test in Alabama but did not conduct one and assumedly came to regret that decision. In *Allen v Alabam* they argued, "As the evidence will show, there was no need to conduct a separate large-scale field tryout in this case, since the first test administration served that purpose" (*Allen*, Defendants' Pre-Trial Memorandum, 1986, p. 113). That decision was unwise because it directly affected the implementation and validity of their procedures. For example, "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" (*Richardson*, 1989, p. 823). In fact, the original settlement in *Allen v Alabama* stipulated that in any new operational examination, the items "shall be field tested using a large scale field test" (*Allen*, Consent Decree, Oct. 25, 1985, p. 3).

The first two administrations of the MECT would have served an important purpose as a full-scale field trial for the new tests, thus avoiding the mistake made in Alabama. However, that opportunity to detect and correct problems in administration, scoring, and interpretation was lost. The impact of the lack of a field trial is further magnified when it is noted that the time period between when NES was awarded the Massachusetts contract (October 1997) and when the first tests were administered (April 1998) was even smaller than the time period NES had to develop the tests in Alabama—a time frame that the court referred to as "quite short" (*Richardson*, 1989, p. 817). Furthermore, even though NES may have drawn many of the MECT items from existing test item banks, items written and used elsewhere still must be field tested on each new population of teacher candidates.

Point-biserial correlations

In the NES *Technical Report* Volume I, Chapter 8, p. 140, there is a description of when an item is flagged for further scrutiny. One of the conditions is when an item displays an "item-to-test point-biserial correlation less than 0.10 (if the percent of examinees who selected the correct response is less than 50)". After such an item is found, "The accuracy of each flagged item is reverified before examinees are scored." The *Technical Report*, however, does not report or provide the percent of persons who selected the correct response on each item. Nor is there an explanation of what the reverification process consisted of, nor of how many items were flagged, nor what was subsequently modified on flagged items. Thus, there is no way to determine the extent to which NES actually followed its own stated guidelines and procedures in the development of the MECT. The relevance of what NES states as their review procedures and what they actually performed is that in Alabama, under the topic of content validity, it was argued by the defense that items rated as "content invalid" were revised by NES and that these "revisions were approved by Alabama panelists before they appeared on a

test." The court, however, found that "no such process occurred" (*Richardson*, 1989, p. 822).

The following table summarizes the point-biserial estimates reported for the MECT. Note that these are not the results prior to NES conducting the item review process. These are the results for the "scorable items" *after* the NES review.

Table 1
Problematic Point Biserial Correlations from the 1999 MECT Technical Report

Date	Number tested	N of M/C Items	Items with point biserials <=0.20					% of total items
			<.00	.00-.05	.06-.10	.11-.15	.16-.20	
Apr-98	4891	315	1	7	15	24	46	29.5%
Jul-98	5716	443	0	2	14	17	39	16.3%
Oct-98	5286	379	2	5	10	15	32	16.9%
Jan-99	9471	507	1	4	14	35	49	20.3%
	25,364	1,644	4	18	53	91	166	332/1644 = 20.2%

Test	Number tested	N of M/C Items	Items with point biserials <=0.20					% of total items
			<.00	.00-.05	.06-.10	.11-.15	.16-.20	
Writing	9750	92	0	0	0	1	1	2.2%
Reading	9455	144	0	0	1	1	6	5.6%
Early Childhood	936	256	0	3	18	30	46	37.9%
Elementary	3125	256	0	2	0	3	27	12.5%
Social Studies	259	128	1	0	1	6	14	17.2%
History	108	64	0	0	2	6	5	20.3%
English	695	256	0	3	11	12	29	21.5%
Mathematics	345	192	1	0	4	4	7	8.3%
Special Needs	691	256	2	10	16	28	31	34.0%
		1,644	4	18	53	91	166	

Source: Massachusetts Educator Certification Tests: Technical Report, 1999

A number of observations may be made from the information in this table. First, of the 1644 total number of items administered over the first four dates, 332 items (20.19%) had point-biserial correlations that are lower than the industry minimum standard criterion of .20. That is a huge percent of poorly performing items for a high-stakes examination. Second, while there are relatively few suspect items on the Reading and Writing tests, there are large numbers of items with poor statistics on many of the subject

matter tests. The Early Childhood, English, and Special Needs tests, in particular, consisted of extraordinarily large percentages of poorly performing items (37.9%, 21.5%, and 34%, respectively). Overall, of the 332 items with low point-biserials, 322 (97%) occurred on the subject matter tests. On the face of it, the results for the subject matter tests are terrible. There is, unfortunately, no authoritative source in the literature (including the *Standards*) that tells us unequivocally whether or not this overall 20.19% of poorly performing items on a licensure examination with high-stakes consequences is acceptable, not acceptable, or even terrible. Given the steps that NES claims were followed in selecting items from existing item banks and in writing new items, there simply should not be this many technically poor items on these tests.

Reliability

In Volume I, Chapter 9, p. 188 of the Technical Report, the following statement appears. “It is further generally agreed that reliability estimates lower than .70 may call for the exercise of considerable caution.” The practical significance of this statement lies in the fact that when reliability is less than .70, it means that at least 30% of the variance in an examinee's test score is attributable to something other than the subject matter that is being tested. In other words, an examinee's test score consists of less than 70% true-score variance and more than 30% error variance. This ratio of true-score variance to error-variance is not desirable in high-stakes examinations (Haney, et al., 1999). Nearly 40 years ago, Nunnally went so far as to describe as “frightening” the extent to which measurement error is present in high-stakes examinations even with reliability estimates of .90 (1967, p. 226).

NES, however, suggests that their reported item statistics and reliability estimates should not greatly influence one's judgment about the overall quality of the tests because the multiple-choice items make up only part of the exam format (NES, 1999, p. 189). The problem with that argument, as noted by Judge Thompson in *Richardson* (1989, pp. 824-25), is that small errors do accumulate and can invalidate the use for which the test was developed. This issue of simply dismissing troubling statistics as inconsequential is particularly ironic when the MECT has been described by the non-profit Education Trust as “the best [teacher test] in the country” (Daley, Vigue & Zernike, 1999).

The Special Needs test deserves closer attention because it had problems at each reported administration.

1. The sample sizes for the tests were 131, 206, 154, and 200, respectively. Based on NES's own criteria (NES, 1999, p. 187), these sample sizes are sufficient for the generation of statistical estimates that would be relatively unaffected by sampling error.
2. The KR-20 reliability coefficients for the four administrations were .67, .76, .76, and .74, respectively. These are minimally tolerable for the last three administrations. The reliability is not acceptable, however, for the first administration. This means that people were denied certification in Special Needs based on their performance on a test that was deficient even by NES's own guidelines.
3. For the April 1998 administration eleven Special Needs items had point-biserials of .10 or less (again, one of NES's stated criterion for “flagging” an item). For the July 1998 administration it was five items, for October 1998 it was four items, and for January 1999 it was eight items. In fact, in two of the administrations there was an item with a negative point-biserial. (Given the previous discussion about the way

the point-biserials were likely to have been calculated (uncorrected), the frequency of negative point-biserials would likely increase if the corrected coefficients had been reported.) Given that there is no specific information about flagging, deleting or replacing items, it is possible that these same faulty items were, and continue to be, carried over from one administration to the next.

The Linkage between Alabama and Massachusetts: A *modus operandi*

At this point the reasonable reader might ask why I am expending so much effort upon what appears to be a relatively minor problem—some items had negative point-biserial correlations. NES, for example, would likely call this analysis "item-bashing", as this type of analysis was referred to in Alabama. The significance of these findings lies in the apparent connection between NES's work in Alabama and their present work on the MECT in Massachusetts.

In Alabama, defendants claimed that

> Before any item was allowed to contribute to a candidate's score, and before the final 100 scorable items were selected, the item statistics for all the items of the test were reviewed and any items identified as questionable were checked for content and a decision was made about each such item (*Allen*, Defendants' Pre-Trial Memorandum, 1986, pp. 113-14).

In fact, in Alabama there were negative point-biserial correlations in the original reliability reports generated by NES (their own documents reported negative point-biserial correlations as large as -0.70) and those negative point- biserial correlations for the same scorable items remained after multiple administrations of the examinations. Simply taking out the worst 20 items in each test did not remove all the faulty items since each exam had to have 100 scorable items. As seen above in Table 1, the MECT has statistically flawed items on many tests, these items have been there since the first administration, and they may be the same items still being used in current administrations.

In Alabama, the negative point-biserial correlations led to the discovery of items for which there was no correct answer. Also discovered were items for which there were multiple correct answers and there were items for objectives that had been rated "not as job related." Additionally, items were found to have been mis-keyed on the item analysis scoring forms. Furthermore, those flawed items existed unchanged for the first eight administrations of the tests. They were not revised, deleted, or changed to "experimental" non-scorable status until the ninth administration--one month after the plaintiffs' team agreed to take the case. Defendants argued that "problems with the testing instrument—such as mis-keyed answers" were simply one component of many that is taken into account by the "error of measurement" (*Allen*, Defendants' Pre-Trial Memorandum, 1986, pp. 108- 113). (Note 10)

As noted earlier, poor item statistics may result for many reasons. Of those reasons the only acceptable one is that they may be due to sampling error (chance). That explanation is unlikely with respect to the MECT, however, because the sample sizes are sufficiently large, and the pattern of faulty item statistics persists over time. The extent to which flawed items may exist in the Massachusetts tests can only be determined by release of the student-level item response data and the content of the actual items, something that has not been done to date. Furthermore, such a release of additional technical information, or item response data, or item content is highly unlikely. (Note 11)

In Alabama, the statistical results and in-house documents were not produced by NES until the plaintiffs seriously discussed contempt of court actions against NES personnel. Consequently, there is little reason to expect that NES will voluntarily release MECT data or results not explicitly covered in their original confidential contract.

In Alabama there were no independent testing experts appointed or contracted to monitor the test developer's work. This fact led the court to conclude that "The developer's work product was accepted by the state largely on the basis of faith" (*Richardson*, 1989, p. 817). In Massachusetts the original MECT contract called for the contractor to recommend a technical review committee of nationally recognized experts who were external to their organization (MDOE, 1997, Task 2.14.i, p. 11). The committee was to review the test items, test administration, and scoring procedures for validity and reliability and was to report its findings to the Department of Education. NES did not form such an independent technical advisory committee for the MECT nor has a formal independent review of the MECT been undertaken by anyone else.

It is not in the short-term business interests of a testing company to conduct disconfirming studies on the technical quality of their commercial product. The MECT is, of course, a product that NES markets as an example of what they can build for other states who might be interested in certification examinations. It is, however, in the best interests of a state for such studies to be conducted. For example, the Commonwealth of Massachusetts has a statutory responsibility to "protect the health, safety and welfare of citizens" who seek services from licensed professionals (NES, 1999, p. 16). In the present situation "citizens" are defined by the Board of Education as "the children in our schools" (MDOE, Special Meeting Minutes, 1998). What has apparently been lost in all of this is the fact that prospective educators are "citizens" and deserve protection too--protection from a faulty product that can damage the profession of teaching and can alter drastically the career paths of individuals. Educators and the public at large deserve the highest quality certification examinations that the industry is capable of providing. There is ample evidence that the MECT may not be such an examination.

Conclusion

A technical review of the psychometric characteristics of the MECT has been called for in this journal (Haney et al. 1999; Wainer, 1999). The year 2000 and 2001 budgets passed by the Legislature of the Commonwealth also called for such an independent audit of the MECT. Those budget provisions, however, were vetoed by Governor Cellucci, and the legislature failed to override the vetoes. Until an independent review committee with full investigative authority is convened by the Commonwealth, the only technical material publicly available for independent analysis is the *1999 MECT Technical Report* generated by NES (NES, 1999). (Note 12) One of the important points made by Haney et al, (1999) was that the Massachusetts Department of Education is not the appropriate agency for conducting such a review. Part of my point here is that the only review of the MECT the Commonwealth may ever see is the one prepared by NES of its own test. Such a review clearly raises a concern over conflict-of-interest (Madaus, 1990; Downing & Haladyna, 1996).

Given the national interest in "higher standards" for achievement and assessment, it must be recognized that there are no "gold" standards by which a testing program such as the MECT can be evaluated (Haney & Madaus, 1990; Haney, 1996). This is ironic given how technically sophisticated the testing profession has become. Consequently, without "gold" standards to define test development practice, there are no legislated penalties for faulty products (tests) and there is no enforced protection for the public. Testing

companies may lose business if the details of shoddy practice are made known and the public may appeal to the judicial system for damages. But the opportunity for a test taker simply to raise a question about a test that can shape his or her career and to have that question taken seriously by an impartial panel should be the right of every test-taking citizen. (Note 13)

Contrary to former Chairman John Silber's statement to the Massachusetts Board of Education, “there is nothing wrong with this test” (Minutes of the Board, Nov. 11, 1998) and the statement by the chief of staff for the MDOE, Alan Safran, “[the test]does not show who will become a great teacher, but it does reliably and validly rule out those who would not” (Associated Press, 1998), there is ample evidence that there may be significant psychometric problems with the MECT. These problems, in turn, have significant practical ramifications for certification candidates and the institutions responsible for their training.

Is the MECT sound enough to support assertions that the candidates are “idiots”? No. Is there evidence that poor performance may, in part, reflect a flawed test containing defective items? Yes. Should the Massachusetts Commissioner of Education independently follow through on the twice-rejected Senate bill to "select a panel of three experts from out-of-state from a list of nationally qualified experts in educational and employment testing, provided by the National Research Council of the National Academy of Sciences, to perform a study of the validity and reliability of the Massachusetts educator certification test as used in the certification of new teachers and as used in the elimination of certification approval of teacher preparation programs and institutions to endorse candidates for teacher certification?" (Massachusetts, 1999, Section 326. (S191K)). Absolutely. Should such a panel serve as a blueprint for the formation of a standing national organization for test review and consumer protection? Yes.

As we enter the 21st century, high stakes tests are becoming increasingly powerful determinants of students' and teachers' lives and life chances. Title II of the 1998 Higher Education Act, in particular, has encouraged a kind of de facto national program of teacher testing. Given the extraordinarily high stakes of these tests, the personal and institutional consequences of poorly designed teacher tests have become too great simply to allow test developers to serve as their own (and lone) quality control and their own (and often non-existent) dispute resolution boards.

Now is the time for the community of professional educators and psychometricians to take a stand and demand that test developers be held accountable for their products in the test marketplace. What this would require at the very least are (1) a mechanism for an independent external audit of the technical characteristics of any test used for high stakes decisions, and (2) a mechanism for the resolution of disputed scores, results, and cases.

Only then will taxpayers, educators, and test candidates have confidence that teacher tests are actually providing the information intended by legislative actions to raise educational standards and enhance teacher quality. Title II legislation certainly did not cause the high stakes test Juggernaut that is rolling through all aspects of educational reform in the U.S. and elsewhere. With mandatory teacher test reporting now tied to federal funding, however, Title II legislation certainly has added to the size, weight, and power of the test Juggernaut and strengthened its hold on reform. For this reason, federal policy makers are now responsible for providing legislative assurances that the public will be protected from the shoddy craftsmanship of some tests and some testing companies and that there will be remedies in place to right the mistakes that result from negligence. This article ends with a call to action. Policy makers must now incorporate into the federal legislation that requires state teacher test reporting new concomitant requirements for the establishment of independent audits and dispute resolution boards.

Notes

I wish to thank Marilyn Cochran-Smith, Walt Haney, Joseph Herlihy, Craig Kowalski, George Madaus, and Diana Pullin for their advice and editorial comments.

1. The class consisted of “all black persons who have been or will be denied any level teaching certificate because of their failure to pass the tests by the Alabama Initial Teacher Certification Testing Program.” (Order On Pretrial Hearing, 1984).
2. This specific wording does not appear until the Amended Consent Decree of Jan. 5, 2000.
3. Among other things, conditions were set on the development of new tests, an independent monitoring and oversight panel was established, grade point averages were ordered to be considered in the certification process, and defendants would pay compensatory damages to the plaintiffs and plaintiffs' attorneys' fees and costs (Consent Decree, 1985).
4. That decision has been upheld numerous times since. The latest Amended Consent Decree was approved on January 5, 2000 (*Allen*, Jan.5 , 2000).
5. George Madaus, Joseph Pedulla, John Poggio, Lloyd Bond, Ayres D'Costa, Larry Ludlow.
6. “Failure tables” consisted of an applicant's name, their raw scores on the exams, the exam cut-scores, their actual responses to suspect items, and their recomputed raw scores if they should have been credited with a correct response to a suspect item. Examinees were identified in court who had failed an examination by one point (i.e., missed the cut- score by one item) but had actually responded correctly to a miskeyed item. For example, on the fifth administration of the Elementary Education exam there were six people who should have been scored correct on scorable item #43 (the so-called “carrot” item) but were not. Their total scores were 72. The cut-score was 73. These individuals should have passed the examination. There was even a candidate who took an exam multiple times and failed but who should have passed on each occasion.
7. The standard contract for test development will include some specification of indemnification. In the case of a state agency like the MDOE, the Request For Responses will typically specify protection for the state, holding the contractor responsible for damages (MDOE, 1997, V. (G), 1, p.17). Contractors, understandably, are reluctant to enter into such an agreement and have been successful in striking this language from the contract.
8. The rationale is that .20 is the minimum correlation required to achieve statistical significance at alpha=.05 for a sample size of 100. This is because .20 is twice the standard error (based on a sample of 100) needed to differ significantly from a correlation of zero.
9. The difference between piloting test items, as NES did, and conducting a field-trial is that the field-trial simulates the actual operational test-taking conditions. Its value is that problems can be detected that are otherwise difficult to uncover. For example, non-standardized testing conditions created numerous sources of measurement error on the first administration of the MECT (Haney et al, 1999).
10. This interpretation of measurement error goes considerably beyond conventional practice where “Errors of measurement are generally viewed as random and unpredictable.” (*Standards*, 1999, p. 26). A miskeyed answer key is not a random error. It is a mistake and its effect is felt greatest by those near the cut-score.

Although false-positive passes may benefit from the mistake, it is the false-negative fails who suffer and, as a consequence, seek a legal remedy.

11. To date the MDOE has routinely ignored questions requesting technical information, e.g. how many items originally came from item banks, who developed the item banks, how many items have been replaced, what are the reliabilities of new items, what are the technical characteristics of the present tests, will the Technical Report be updated, what "disparate impact" analyses have been conducted?
12. From the start of testing to the present time individual IHE's have not been able to initiate any systematic analysis of their own student summary scores, let alone any statewide reliability and validity analyses. The primary reason for this paucity of within- and across- institution analysis is because NES only provides IHEs with student summary scores printed on paper—no electronic medium is provided for accessing and using one's own institutional data. Thus, each IHE faces the formidable task of hand-entering each set of scores for each student for each test date. This results in a unique and incompatible database for each of the Commonwealth's IHEs.
13. I assert that the right to question any aspect of a high-stakes examination should take precedence over the waiver required when one takes the MECT: "I waive rights to all further claims, specifically including, but not limited to, claims for negligence arising out of any acts or omissions of the Massachusetts Department of Education and the Contractor for the Massachusetts Educator Certification Tests (including their respective employees, agents, and contractors)" (MDOE, 2001, p. 28).

References

Angoff, W. (ed.). (1971). *The College Board Admissions Testing Program: A Technical Report on Research and Development Activities Relating to the Scholastic Aptitude Test and Achievement Tests*. NY: College Entrance Examination Board.

Allen v. Alabama State Board of Education, 612 F. Supp. 1046 (M.D. Ala. 1985).

Allen v. Alabama State Board of Education, 636 F. Supp. 64 (M.D. Ala. Feb. 5, 1986).

Allen v. Alabama State Board of Education, 816 F. 2d 575 (11th Cir. April 22, 1987).

Allen v. Alabama State Board of Education, 976 F. Supp. 1410 (M.D. Ala. Sept. 8, 1997).

Allen v. Alabama State Board of Education, 190 F.R.D. 602 (M.D. Ala. Jan. 5, 2000).

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, D.C.: American Educational Research Association.

Associated Press Archives, (October 4, 1998). *State Administers Teacher Certification Test Amid Ongoing Complaints*.

Baldus, D.C. & Cole, J.W.L. (1980). *Statistical Proof of Discrimination*. NY: McGraw-Hill.

Cochran-Smith, M. (in press). The outcomes question in teacher education. *Teaching and Teacher Education*.

Cochran-Smith, M. & Dudley-Marling, C. (in press). The flunk heard round the world. *Teaching Education*.

Consent Decree, *Allen v. Alabama State Board of Education*, No. 81-697-N (M.D. Ala. Oct. 25, 1985).

Crocker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. NY: Holt, Rinehart and Winston.

Daley, B. (1999). "Teacher exam authors put to the test". Boston Globe, 10/7/98, B3.

Daley, B.; Vigue, D.I. & Zernike, K. (1999) "Survey says Massachusetts Teacher Test is best in US". *Boston Globe*, 6/22/99, B02.

Defendant's Pre-trial Memorandum, *Allen v. Alabama State Board of Education*, No. 81-697-N (M.D. Ala. May 1, 1986).

Donlon, T. (ed.) (1984). *The College Board Technical Handbook for the Scholastic Aptitude Test and Achievement Tests*. NY: College Entrance Examination Board.

Downing, S. & Haladyna, A. (1996). A model for evaluating high stakes testing programs: Why the fox should not guard the chicken coop. *Educational Measurement: Issues and Practice, 15:1*, pp.5-12.

Dressel, P.L. (1940). Some remarks on the Kuder-Richardson reliability coefficient. *Psychometrika,* 5, 305-310.

Ebel, R.L. & Frisbie, D.A. (1991) (5th ed.). *Essentials of Educational Measurement*. NJ: Prentice Hall.

Guilford, J.P. (1936) (1st ed.). *Psychometric Methods*. NY: McGraw-Hill.

Guilford, J.P. (1954) (2nd ed.). *Psychometric Methods*. NY: McGraw-Hill.

Haney, W., & Madaus, G. F. (1990). Evolution of Ethical and Technical standards. In R.K. Hamilton, & J. N. Zaal (Eds.), *Advances in Educational and Psychological Testing* (pp.395-425).

Haney, W.M., Madaus, G.F. & Lyons, R. (1993). *The Fractured Marketplace for Standardized Testing*. Boston: Kluwer.

Haney, W. (1996). Standards, Schmandards: The need for bringing test standards to bear on assessment practice. Paper presented at the annual meeting of the American Educational Research association annual meeting. NY: NY.

Haney, W., Fowler, C., Wheelock, A, Bebell, D. & Malec, N. (1999). Less truth than error?: An independent study of the Massachusetts Teacher Tests. *Education Policy Analysis Archives, 7(4).* Available online at http://epaa.asu.edu/epaa/v7n4/.

Henrysson, S. (1963). Correction for item-total correlations in item analysis. *Psychometrika, 28*, 211-218.

Hopkins, K.D. (1998) (8th ed.). *Educational and Psychological Measurement and Evaluation*. Boston: Allyn and Bacon.

Kuder, G.F. & Richardson, M.W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151-160.

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin, 51*, 493-504.

Lord, F.M. & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Madaus, G. (May 19-20, 1986). Testimony in *Allen v Alabama* (81-697-N).

Madaus, G. (1990). Legal and professional issues in teacher certification testing: A psychometric snark hunt. In J.V. Mitchell, S. Wise, & B. Plake (Ed.), *Assessment of teaching: Purposes, practices, and implications for the profession*. (pp. 209-260). Hillside, NJ: Lawrence Erlbaum Associates..

Massachusetts. (1999). *FY 2000-2001 Budget*.

Massachusetts Department of Education (February 24, 1997). Massachusetts Teacher Certification Tests of Communication and Literacy Skills and Subject Matter Knowledge: Request for Responses (RFR).

Massachusetts Department of Education (July 1, 1998). Board of Education Special Meeting Minutes. *http://www.doe.mass.edu/boe/minutes/98/min07 0198.html.*

Massachusetts Department of Education (July 27, 1999). Department of Education Press Release. *http://www.doe.mass.edu/news/archive99/pr072 799.html.*

Massachusetts Department of Education (November 28, 2000). Board of Education Regular Meeting Minutes. *http://www.doe.mass.edu/boe/minutes/00/1128r eg.pdf.*

Massachusetts Department of Education (February 16, 2001). Massachusetts Educator Certification Tests: Registration Bulletin. *http://www.doe.mass.edu/teachertest/bulletin 00/00bulletin.pdf*

Melnick, S. & Pullin, D. (1999, April). *Teacher education & testing in Massachusetts: The issues, the facts, and conclusions for institutions of higher education*. Boston: Association of Independent Colleges and Universities of Massachusetts.

Millman, J. (June 17, 1986). Testimony in *Allen v Alabama* (81-697-N).

National Board on Educational Testing & Public Policy. (2000). *Policy statement*. Chestnut Hill, MA: Lynch School of Education, Boston College.

National Commission on Testing and Public Policy. (1990). *From Gatekeeper to Gateway: Transforming Testing in America*. Chestnut Hill, MA: Lynch School of

Education, Boston College.

National Evaluation Systems. (1999). *Massachusetts Educator Certification Tests Technical Report*. Amherst, MA: National Evaluation Systems.

Nunnally, J. (1967). *Psychometric Theory*. NY: McGraw- Hill.

Order On Pretrial Hearing, *Allen v. Alabama State Board of Education*, No. 81-697-N (M.D. Ala. Dec. 19, 1984).

Pearson, K. (1909). On a new method of determining correlation between a measured character A and a character B, of which only the percentage of cases wherein B exceeds or falls short of a given intensity is recorded for each grade of A. *Biometrika, Vol. VII*.

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

Rawls, P. (2000). “ACT may design test for Alabama's future teachers.” *The Associated Press*, 7/11/00

Richardson v. Lamar County Board of Education, 729 F. Supp. 806. (M.D. Ala 1989) *aff'd*, 935 F. 2d 1240 (11th Cir. 1991).

Richardson, M.W. & Stalnaker, J.M. (1933). A note on the use of bi-serial r in test research. *Journal of General Psychology, 8*, 463-465.

Thorndike, E.L., Bregman, M.V., Cobb, Woodyard, E. et al., (1929) *The Measurement of Intelligence*. NY: Teachers College, Columbia University.

U.S. Department of Education, National Center for Education Statistics. *Reference and Reporting Guide for Preparing State and Institutional Reports on the Quality of Teacher Preparation: Title II, Higher Education Act*, NCES 2000- 089. Washington, DC: 2000.

Wainer, H. (1999). Some comments on the Ad Hoc Committee's critique of the Massachusetts Teacher Tests. *Education Policy Analysis Archives, 7(5).* Available online at http://epaa.asu.edu/epaa/v7n5.html.

Walden, J.C. & Deaton, W.L. (1988). Alabama's teacher certification test fails. 42 *Ed. Law Rep.1*

About the Author

Larry H. Ludow
Associate Professor
Boston College
Lynch School of Education
Educational Research, Measurement, and Evaluation Department

Email: Ludlow@bc.edu

Larry Ludlow is an Associate Professor in the Lynch School of Education at Boston College. He teaches courses in research methods, statistics, and psychometrics. His

research interests include teacher testing, faculty evaluations, applied psychometrics, and the history of statistics.

The World Wide Web address for the *Education Policy Analysis Archives* is **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 Commentary Editor is Casey D. Cobb: casey.cobb@unh.edu .

EPAA Editorial Board