Protecting the Integrity of Computer-Adaptive Tests:

Results of a Legal Challenge

Paper Presented at the Annual Meeting of the

American Educational Research Association

by:

Gregory J. Cizek

Professor of Educational Measurement and Evaluation

University of North Carolina at Chapel Hill

cizek@unc.edu

April 13, 2004

ABSTRACT

This paper describes the history, perspectives, and outcome of a legal challenge to score invalidation by a group of examinees who were administered a medical specialty in-training examination in computerized mastery testing (CMT) format in 2002. Upon receipt of information which raised concern that some examinees may have had inappropriate prior access to test items, the testing company and specialty board conducted an investigation and subsequently invalidated scores for all test takers from examinees' institution. Examinees whose scores had been invalidated filed suit in federal court alleging that the score invalidations were improper. A trial resulted in a ruling that the specialty board and its testing contractor erred in invalidating the scores. This paper provides a perspective and analysis of this important case. Recommendations are offered to help those who administer tests better prepare for secure test administrations and become more aware of important aspects of validity that can come into play when invalidation of test scores is challenged.

Protecting the Integrity of Computer-Adaptive Tests:

Results of a Legal Challenge[1]

Introduction

More and more commonly, threats to the integrity and validity of testing are being witnessed, particularly as the stakes associated with passing or failing a test increase. Several recent publications have demonstrated that cheating by test takers is on the rise (McCabe & Trevino, 1996) and have suggested strategies for detecting, preventing, and responding to cheating in both large-scale and classroom contexts (Cizek, 1999, 2003). A new company on the testing scene, Caveon, has recently begun a business focused exclusively on providing

> "comprehensive services to combat test fraud and piracy with detection services to identify security breaches, remediation services to confirm suspicions and leverage legal processes to halt abuses of sensitive test information, and prevention services to secure tests from compromise" (2004, p. 1).

The rise in challenges to the integrity of tests has also been accompanied by a rise in legal challenges to testing. Referring to the unique context of licensure and certification testing, Carson (1999) has stated: "Given the consequences, it is fair to assume that legal challenges will continue" (p. 442). In the licensure and certification arena, legal challenges take on particular importance. These challenges often involve differing and sometimes competing interests; namely, the rights and responsibilities of

individual test takers, the interests of those responsible for tests in developing and administering examinations that yield accurate inferences about test takers; and state and professional association responsibilities for protecting the public from incompetent or harmful practitioners.

Background

The legal challenge analyzed in this paper concerns a cohort of students in a medical specialty whose scores on an in-training examination had been invalidated by the specialty board overseeing the examination and licensure process. The test construction, administration, and scoring involved a form of computerized mastery testing. The following subsections provide some background on the testing approach and on the context for the case.

Computerized Mastery Testing

The test design and administration format involved an adaptive algorithm called computerized mastery testing (CMT) described elsewhere (see, e.g., Gibley, 1998; Lewis, & Sheehan, 1988). In the case of the NBPME examinations, examinees were administered a series of testlets (Lewis & Wainer, 1990). Testlets are developed to the same target set of specifications so that they are equivalent in content coverage, difficulty, and reliability.

For the NBPME examinations a moderately large pool of possible testlets exited. However, the testlets administered to an examinee are selected via an algorithm from a specified (smaller) pool of testlets that are operationalized for a given administration. Each testlet for the NBPME examinations consisted of 15 items. An examinee could be administered a minimum of 4 testlets; a maximum of 10 testlets could be administered. At any point between the administration of 4 and 10 testlets, inclusive, if sufficient information has been obtained to make a pass/fail decision, the test administration would be complete.

Score Invalidations

Examinees (n = 33) were first-year podiatry students at the Ohio College of Podiatric Medicine (OCPM). The students took a computer-administered version of Part I of the National Board of Podiatric Medical Examiners (NBPME) examination during a "window" for taking the computerized test that spanned July 9-12, 2002. The case described here involves a defendant medical specialty board and its testing contractor, the Chauncey Group International, and the group of plaintiff students and their institution, OCPM.

The origin of the current legal challenge case can be traced to certain testing irregularities and other information that came to light following the July 2002 test administration. Following the administration of that test, the testing company that developed and administered the NBPME examinations received information which raised concern that some examinees may have had inappropriate prior access to test items. The testing contractor relayed this information to the NBPME.

On August 13, 2002, NBPME sent a letter to the OCPM examinees indicating that their test scores would be delayed due to a testing irregularity. Following its investigation of the matter, Chauncey recommended to NBPME that scores for all OCPM students who had taken the July 2002 examination be invalidated. On October 21, 2002, NBPME sent a letter to the July 2002 examinees from OCPM informing them that their scores were invalidated.

In January 2003, examinees whose scores had been invalidated filed suit in federal district court alleging that the score invalidations were improper; that NBPME/Chauncey had engaged in deceptive trade practices, had breached a contract with examinees, and had defamed examinees. A jury trial occurred over a period of approximately six weeks (Ohio College of Podiatric Medicine [OCPM], et al., v. National Board of Podiatric Medical Examiners [NBPME], et al, 2003). On April 8, 2003 Judge Donald Nugent "ruled that the National Board of Podiatric Medical Examiners and its outside exam developer erred when they threw out the test scores...based on an unfounded suspicion of cheating" (Elrick, 2003, p. 1).

Aspects of the Legal Challenge Involving Validity Concerns

Of the three bases for plaintiffs' complaint, two were upheld. The jury found that the defendants had breached a contract with plaintiffs and that defendants had engaged in unfair trade practices. The jury found that the publication of information related to defendants' actions and score invalidations was not defamatory. The two causes that the jury upheld, however, suggest avenues for improving the validation efforts for tests generally, and improvements in procedures that can be followed when score invalidations are contemplated. Four aspects related to these improvements are described in the following sections.

"The Window"

Although computerized test administrations have been conducted for many years now, testing professionals continue to wrestle with technological aspects ranging from those that are purely practical to those that affect score validity. One example of the former can be seen in a recent experiment in the state of North Carolina. The state developed and implemented a large-scale, computer-adaptive, end-of-course tests in reading and mathematics for a subpopulation of students in grades 3 through 8. In the first year the system was tried out, as soon as the test became available for administration in the field, the number of simultaneous test administration requests quickly overwhelmed the state's technology infrastructure and the testing schedule was disrupted.

It is perhaps these kinds of potential practical problems and logistical constraints that have been taken into consideration in the scheduling of other, large-scale CAT or CBT administrations. One scheduling accommodation is sometimes referred to as a test *window*. The window is a specified period of time that an operational test form will be made available to examinees at a given test site. For example, suppose that a credentialing body wished to administer its examination in CBT on a national scale. Further suppose that the credentialing body contracted with a separate company for space at test

sites and test delivery services. Finally, suppose that the examination would require approximately four to five hours of examinee time, that a typical test site could accommodate up to 15 examinees at a single time, and up to 30 examinees per day. If the credentialing body has a fairly large number of potential test takers, and/or if the number of test sites is limited, it is possible that not all examinees could be tested on a single day. (This is particularly a possibility if, as is typical, the test site/test delivery provider contracts with other credentialing agencies or test providers for services.)

To address this capacity issue, a testing "window" is sometimes created. A testing window can be defined as a specified number of consecutive days during which a particular test will be available at a testing site. The availability of a testing window may permit more convenient scheduling for all examinees, and permits the testing center to be utilized in an efficient manner. Unfortunately, a testing window also creates conditions that can result in serious threats to the validity of scores.

The presence of a testing window was a relevant, though indirect, factor in the legal challenge described in this paper. The case of the OCPM score invalidations was part of a larger investigation into possible inappropriate activities involving other colleges. At one college in another state, information was reported indicated that the use of a testing window was exploited by some students, apparently in a collaborative effort to obtain inaccurate test scores. The scheme consisted of some highly able students at that location registering and taking the NBPME examination early in the testing window. At least some of those students would then apparently record and transmit information about the test to other students who would take the examination later in the testing window. The pattern of test scores and registration volumes for that test site provided confirming evidence of the report.

Apparently, the NBPME Part I examination had been administered in a four-day window since at least 1992. In that year, a similar concern about preknowledge (in several states) was brought to the attention of the credentialing organization. A security concern was raised involving test takers taking the test early in the testing window and passing along information to examinees who took the test later in the testing window. According to the testimony of one of the witnesses for the testing company, the current

situation represented the "same test, same basic situation where information was available to students in advance" as was present in 1992.[2] As a result of the breach in 1992, scores for examinees from six of seven podiatry schools were invalidated.

The fact that the testing window problem was so serious--and that the credentialing organization and testing company had been aware of the problem for 10 years--is cause for concern. The *Standards for Educational and Psychological Testing* require that "reasonable efforts should be made to assure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent means" (AERA/APA/NCME, 1999, p. 64). Similarly, Standard 10.5 of the *ETS Standards for Quality and Fairness* in testing requires that test administrations be conducted in such a way as to "eliminate opportunities for test takers to attain scores or other assessment results by fraudulent means, to the extent possible" (ETS, 2000, p. 51).

In the current case involving OCPM students, it was not alleged that they gained unfair advantage by or had used the "window scheme." Nonetheless, information obtained related to the scheme in the other state was relevant for several reasons. One prominent reason is relevant to this discussion of test security. Technology makes the rapid and wide dissemination of information possible. Thus, inappropriate access to, recording of, and dissemination of test content in one location could easily be transmitted to other test sites and potential examinees in far-removed locations. There only known way to prevent this threat to validity of scores in a single location (or across multiple locations) is to make the testing window as narrow as possible, even to the point narrowing it to a single day.

Disconfirming Evidence

In the testing field, the current, professionally-accepted conceptualizations of validation are often expressed as metaphor that involves the gathering of evidence or arguments that bear on some intended meaning or interpretation of a test score (see, e.g., Kane, 1992; Messick, 1989). A key aspect of current validation theory and practice is the recognition that *inference* is required to interpret any test score or

examinee performance.

According to the principle of validation requiring inference, validation is a process by which those responsible for the development, administration, and use of tests engage in "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores..." (Messick, 1989, p. 13, emphasis in original). That is, inference is always required to interpret any test score, and evidence must be gathered related to the intended interpretation(s) of the scores.

Professionally-defensible validation requires "ascertain[ing] the degree to which multiple lines of evidence are consonant with the [intended] inference, *while establishing that alternative inferences are less well supported*" (Messick, 1989, p. 13, emphasis added). In the testing profession, this requirement has sometimes been referred to as a requirement to gather both "confirming" and "disconfirming" evidence bearing on interpretation of a test score. That is, those responsible for testing programs have an obligation to gather evidence that the test scores mean what they are intended to mean (*confirming evidence)*, as well as evidence that test scores do not mean something other than that which is intended *(disconfirming evidence)*.

There is good reason for searching out both kinds of evidence to the extent practical. Such an approach helps account for bias or "blind spots" in the process of validation. In his chapter on validity, Messick (citing Campbell, 1960) provides:

> an important cautionary note: This very variety of methodological approaches in the validational armamentarium, in the absence of specific criteria for choosing among them, *makes it possible to select evidence opportunistically and to ignore negative findings* (1989, p. 33, emphasis added).

Messick (referring to the work of Kaplan, 1964) elaborated on his caution by reminding testing professionals that:

"Values are important to take into account in score interpretation not only because they can

directly bias score-based inferences and actions, but because they could also indirectly influence in more subtle and insidious ways the meanings and implications attributed to test scores..." and he urged explicit recognition of the potential for "adherence to values of such a kind and in such a way as to interfere with scientific objectivity" (1989, p. 59)

Messick concluded that:

Everything depends on the way in which the [validation] inquiry is conducted and the conclusions derived..... [S]cience does not demand that bias be eliminated but only that our judgment take it into account. (1989, p. 59).

Thus, while an effort to gather both confirming and disconfirming evidence does not completely eliminate inferential blind spots, the gathering of on ly evidence that supports a preferred inference can operate to immunize the intended inference against plausible, rival interpretations.

The written record in this legal challenge did not reveal any indication that the NBPME board or the testing contractor sought out or considered any disconfirming evidence. It should be noted here that this does not mean that the board or the testing personnel did not do so--only that if such an activity took place, it was not recorded. (A subsequent section of this paper will deal more directly with documentation.) Suffice it to say that there was a lack of evidence that disconfirming evidence was explored or alternative inferences were considered.

In fact, to the contrary. The record indicated that the board and testing company appeared to consider only a single inference; namely that the scores of OCPM students potentially reflected inappropriate preknowledge of test information. A weak counter-argument may be that vigorous collection and evaluation of disconfirming evidence is not the norm. That may be unfortunately true. Indeed one long-acknowledged flaw in test validation efforts is that they are not as searching and enthusiastic as they ought to be. As Robert Ebel wrote more than 40 years ago:

"Validity has long been on of the major deities in the pantheon of the psychometrician. It is universally praised, but the good works done in its name are remarkably few." (1961. p. 640)

However, it should be obvious that the fact effort in an area has been recognized as weak for decades must not be confused with an endorsement of weakness or with codifying the status quo as a standard.

In my own work, I have strongly advocated for honesty and integrity in examinee test performance. I know that this concern for security in test administration is widely endorsed by those who sponsor licensure and certification programs. My work in the field of licensure and certification testing has also made me recognize the great efforts of credentialing bodies with respect to their role in helping to protect the public from unqualified or unsafe practice of a profession. The vigor with which many such organizations pursue these goals of integrity, security, and public protection is essential and admirable.

However, in line with Messick's caution, it is possible that "adherence to values of such a kind and in such a way as to interfere with scientific objectivity" may have resulted in "blind spots" in the process of investigating and pursuing score invalidations in this case described here. Specifically, adherence to a single interpretation (i.e., that cheating had occurred) may have resulted in the failure of the board and its testing company failing to seek any disconfirming evidence (i.e., evidence that supported the interpretation that cheating had not occurred) and, when disconfirming evidence was available, may have resulted in the failure to recognize it as such.

Numerous instances of this problem came to light during the course of the trial. A few, selected examples of this are described in the following sections.

A "Study Guide?"

During the trial, one of the major sources of concern about inappropriate prior access to test material related to a what came to be referred to as a "study guide." The guide contained approximately 1078 entries typed up into a 35-page booklet. Each page had two columns for each entry; the first column

listed a topic or question, the second gave the correct answer. Table 1 gives an example of this format.

Insert Table 1 about here.

As the table shows, the examination review material was consistent with the purpose of the examination; that is, to test candidate's knowledge of anatomy, pharmacology, and basic science facts and principles. An analysis conducted by the testing contractor found that "approximately 202" of the 1078 entries were similar to items found in the NBPME item pool. Apparently, a working hypothesis that drove the testing company's (in)validation efforts was that the study guide was compiled as a result of inappropriate access to ("harvesting" of) the NBPME item pool. It is interesting to consider that the result of the analysis could be equally or even more plausibly interpreted as evidence *disconfirming* the inference that cheating had occurred. For example, the fact that the only a small proportion (18.7%) of those topics could be linked as similar to items in the pool would suggest that the source of topics for the items was *not* from compromised tests but from some other source or sources. Additionally, with the exception of two entries, the entries in the study guide did contain suggested incorrect options (i.e., they were not "items" in the sense that testing specialists use the term); some of the "answers" suggested in the second column of the document were actually incorrect; and the entries that the testing company flagged as "similar to" actual test items were not flagged, highlighted, or in any way distinguishable from the other entries.

Testing Volume

As noted previously, in another state a pattern of testing volume increases across the testing window was observed. In documentation reviewed during the Ohio case, NPBME/Chauncey considered

patterns of "testing volume" to be a "piece of the puzzle" used to arrive at the inference of inappropriate preknowledge. It seems easy to concur that this evidence is relevant and supports an inference of cheating. In the OCPM case, however, it was found that a smaller proportion of OCPM students tested later in the window, and that the largest proportion of students tested on the first day the test was made available. Logically, if a pattern of testing volume increases across the testing window is considered to be evidence confirming an inference of cheating, it stands to reason that the opposite pattern of testing volume should be considered as evidence disconfirming an inference of unfair advantage. However, the record and testimony presented at the trial did not reveal that the credentialing body or the testing contractor recognized or considered this as disconfirming evidence in arriving at the decision to invalidate scores for OCPM examinees.

Examinee-Provided Information

The written record and testimony indicated that the credentialing board and the testing company considered that evidence of score gains for students who took the test on more than one occasion supported an inference of potential preknowledge by OCPM examinees. An analysis of scores for OCPM examinees who had taken the test more than once resulted in 4 examinees considered to be "suspicious repeaters." An investigation, including interviews or attempted interviews with various persons (including OCPM faculty, administrators, and examinees) was pursued.

It should be obvious that, if a student has cheated, her or she may not be forthcoming during an interview and accounts they provide cannot automatically be considered to be truthful. Nonetheless, the accounts provided by those interviewed *can* be investigated for their veracity. One striking example of disconfirming evidence offered by one of the examinees identified as a suspicious repeater is found in the account of Ms. Misty Baker. According to information introduced at the trial, Ms. Baker was interviewed in the course of an investigation. The investigator reported that:

Ms. Baker said she did very well on the July 2002 NBPME Part I and that this is the second type

> [sic] she has taken the test. She advised that she used the USMLE study guide and class notes to prepare for the test. Ms. Baker took the test on the first day and did not talk to anyone. When I advised Ms. Baker of the large score increase from the first time she took the test, she replied that she was pregnant the first time and has an older daughter. For the second time, Ms. Baker said she sent her two children to her parents in Texas and spent 10-12 hours every day preparing for the test.

A reasonable conclusion from this example (and others which appear in the record) is that plausible evidence to disconfirm the inference of inappropriate access was clearly available. Some follow-up may have been required to substantiate the account provided by Ms. Baker but, if corroborated, would seem to be evidence that would weigh on the side of disconfirming and inference of cheating. Unfortunately, from the record for this trial, it appears that such follow-up did not occur. More importantly, there was an absence of evidence to suggest that NBPME/Chauncey recognized or considered what would appear to be highly plausible evidence to contradict the inference of inappropriate preknowledge.

In summary, at some point, a preponderance of evidence --I believe--strongly urges the rejection of one interpretation in favor of an alternative. However, this is only the case if all of the available evidence is weighed in the validation process of confirming or disconfirming the intended inference. Overall, it was my conclusion that obviously disconfirming evidence was available to NBPME/Chauncey, but that these sources of evidence were either unrecognized as such or ignored. The process of investigating the meaning of the OCPM examinees' scores likely suffered from a strong confirmation bias favoring invalidation.

Documentation

Even if defensible approaches are pursued in validation efforts (for example, the good-faith

search for both confirming and disconfirming evidence bearing on an intended interpretation of scores) those efforts may not appear as such in the absence of documentation. Such documentation is, however, not necessarily (or even primarily) desirable solely to accomplish the goal projecting the appearance of an impartial inquiry. Documentation can provide evidence that appropriate procedures and data sources, were used, and that results upon which conclusions rest were accurate.

In legal challenge described in this paper, there were a number of junctures at which the presence of adequate documentation (or, *any* documentation) would have been of great benefit in defending the propriety of score invalidations. A sample of those junctures are described in the following sections.

Keyword Analysis

The case record indicates that an individual from the testing contractor conducted what came to be referred to as a keyword analysis for establishing similarity between items in the NBPME item pool and the study guide. However, no detail was provided regarding the procedures used, and the individual responsible for conducting the analysis may not have been able to accurately recall the specific procedures and decision rules used and other important details. For the sake of illustration, let us assume that the keyword analysis was performed by examining the first entry in the study guide and distilling two or three major concepts or terms from the entry. Then, each term could be used as a search string against the text of items in the item pool. In fact, the search may not necessarily even need to comprise all text of all items in the pool. Because many item banks code each item entered with information such as the correct answer, an item identification code, content classification(s), statistics for the item's performance, and key words describing the focus of the item, it might be that a similarity analysis would only involve terms appearing in a "keyword field."

Herein lies a first difficulty. Even the most basis information related to the analytical procedures was not documented. Thus, for example, we do not know if correspondence between the entries from the study guide and the item pool was conducted based on keyword field information or against the entire text

of an item, or in some other manner.

Relatedly, it is important to recognize that there is no professionally-recommended or described method for conducting such an analysis. Thus, it becomes especially critical to document the actual procedure used in order to evaluate whether the keyword analysis was appropriate, accurate, and reproducible.

According to person responsible for conducting the analysis, the method she used for determining similarity consisted of the following steps and assumption:

> "I think we determined that if you read the question in this document [referring to Exhibit 59], if you memorized this question, and that would allow you to answer our item correctly, we would assume that this item was in our pool...."

Logically, the procedure seems inappropriate on its face. For example, suppose the same procedure were applied to review questions that might appear at the end of a chapter in a textbook in the field of podiatry. According to the procedure used, if a student read and memorized the question such that it allowed the student to answer correctly an item in the NBPME, it would be assumed that the item came from the NBPME item pool. Such an interpretation would lead to the disqualification as "compromised" of any item presented in any source matching content in the NBPME item pool. In effect, such a principle could be taken to mean that the only review questions a student could ethically study would be on topics not covered by the examination. Plainly, such a result is absurd.

Of course, the method described also has some logical appeal. I am not an expert in such analysis, but I if I needed to perform such an analysis, I might propose a keyword search procedure along the lines of that which was hypothesized earlier in this paper. Thus, the point here is *not* that the procedure was inappropriate, simply that its propriety was not documented and, consequently, its propriety, accuracy, and reproducibility cannot be evaluated.

For a moment, let us examine more closely the issue of reproducibility. Let us suppose that the testing company person who conceived of and conducted the keyword analysis used a matching procedure similar to that hypothesized to flag potentially similar items. That is, a key term from the study guide would be entered into the item bank database to search the text of all items in the pool for the one or more items for which might contain the specified term. According to information revealed in the course of the trial, whatever procedure was used yielded matches for approximately 202 items. However, the technical details of how the searches were conducted (if they had been described) would be the only objective aspect of the process used to ascertain similarity. It is almost certain that multiple items from the item pool would be identified as possible "matches" because the each contained the searched-for key word. The critical step of comparing the item in its entirety with the key word from the entry in the study guide is highly subjective. The written record did not indicate the decision rules for making the critical determination.

Further, it would seem that certain qualifications and expertise would be essential to conduct such a comparison between a single key word and entire items from the pool. Such qualifications might include familiarity with test development, training in linguistics, structural analysis, or podiatry among other possible qualification that would suggest confidence in the matching procedure itself or in any conclusions that an entry in the study guide should be counted as "similar" to an item in the pool.

Because no professionally-accepted standard procedure for establishing such similarity exists, it would have been desirable to conduct--and document--the results of repeated implementations of whatever procedure was used. It is almost certain that the number of items counted as "matches" or "similar" that might be identified by another equally qualified, independent reviewer using the same procedure would differ, and the degree of agreement would be an important criterion in assessing the results of the procedure. Finally (though I am not suggesting this), it is possible that an equally intuitively-appealing procedure could be devised. It, too, would almost certainly yield different results and the extent of agreement between equally acceptable procedures would, if substantial, support

whatever inference or evidence was yielded by the procedure actually used.

Statistical Procedures Used to Support an Inference of Cheating

A number of statistical analyses were conducted to flag examinees as potentially having preknowledge or "suspicious" score gains. Of major concern is that the specific statistical tests used were, in many cases, not described.

For example, the case record indicates that an analysis was conducted in which the entire available national population of examinees' performances were compared on two reconfigured tests: one comprised of potentially compromised testlets and another comprised of testlets considered as not likely compromised. However, the statistical procedure used to arrive at conclusions that inappropriate preknowledge existed was not specified. Thus, no independent, qualified psychometrician or statistician would be able to verify the accuracy of any conclusions suggested by those tests. Behind this concern about documentation and replicability lies the principles articulated in the *Standards for Educational and Psychological Testing.* According to the *Standards*:

> In educational testing programs and in licensing and certification applications, when it is deemed necessary to cancel or withhold a test taker's score because of possible testing irregularities, including suspected misconduct, the type of evidence and procedures used to investigate the irregularity should be explained to all test takers whose scores are directly affected by the decision. (AERA/APA/NCME, 1999, p. 89)

As another example of insufficient documentation lies in the statistical analyses performed and which suggested the conclusion of "suspicious" score gains of students who took the NBPME examination on multiple occasions (i.e., repeaters). One of the exhibits made available during the trial showed presented data related to scores obtained by repeat test takers on each occasion on which they

were administered the NBPME Part I examination. Four such repeat test takers are identified from the OCPM. The record *did* contain the testing company's conclusion about these test takers; namely, that they "increased their results so dramatically so as to be classified as suspicious repeaters" and that "four Ohio student were identified as having statistically significant large-score gains."

Unfortunately, the written record did not provide any details about whatever statistical test(s) were performed to arrive at that conclusion. For example, the statistical test used was not named or described. Basic data necessary to conduct/replicate a statistical test were not provided. Ultimately, it appeared that the conclusion that score gains were dramatic appears to have been based solely on "eyeballing" pairs of test scores and concluding that one number is higher than another. In a deposition, the testing company person responsible for conducting such an analysis was asked the following question: "Now tell me what you did to perform this analysis other than what's listed here [in an exhibit], which is to list the prior dates and scores and compare it with the 2002 test score?" The response: "That's essentially it."

In summary, it is not known what impact the lack of specificity regarding procedures or absence of documentation had on the outcome of this case. It may be that defensible procedures were used that may have resulted in accurate and reproducible inferences. On the other hand, the lack of documentation of whatever procedures were used introduces questions and doubt about their accuracy and defensibility. There is no need for these to questions to remain in doubt, however, particularly when simple documentation of procedures would put them to rest.

Contract Language

The final aspect of validity that this case raised was the aspect of fundamental fairness as regards the language contained in informational material provided to candidates. Like many similar licensure and credentialing bodies, the NBPME produces a guide for candidates that contains information about its

testing program. This *Bulletin of Information* (NBPME, 2002) covers, among other topics, eligibility, procedures for registration, fees, content outline for the test, sample questions and so on.

Candidate guides can be construed as a contracts in that they outline the rights and responsibilities of the credentialing organization and test takers. Candidates pay a fee to obtain a service or product described in the candidate guide. More specifically, a candidate guide can be seen as what is called a *contract of adhesion*. This is a term of art in the legal profession and it refers to "a contract (often a signed form) so imbalanced in favor of one party over the other that there is a strong implication it was not freely bargained" (Law.com, 2004, p.1). Contracts of adhesion are essentially drafted with input by only one of the parties (i.e., with little or not opportunity for the other party to have input as regards the terms of the contract). Moreover, they are often offered on a "take it or leave it" basis by the party that has the greater role in defining its terms. When challenged, contracts of adhesion typically receive a high level of scrutiny by the courts.

Because the language of the contract will likely receive a high level of scrutiny in the case of a challenge to a score invalidation, it is imperative that the language be crafted carefully and with attention to fundamental fairness--which is clearly an aspect of validity. The *Bulletin of Information* in force at the time of the 2002 examinations contained the following language related to score invalidation and appeal:

> If a candidate's scores are withheld or canceled, that candidate may, within 15 business days of the notification, submit a written request for a hearing. The purpose of the hearing will be to determine whether there exists sufficient, competent, and credible evidence that the candidate acted improperly at the time of the National Board examinations... At the hearing, the candidate may present such evidence as he or she deems proper and necessary. (2002, p. 8)

The *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999) also suggest that permitting input by an examinee whose score has been invalidated is sound professional

practice. According to *Standards*, "Test takers should be given a timely opportunity to provided evidence that the score should not be canceled or withheld" (p. 89).

A reading of the preceding passages would suggest that the plain language of the *Bulletin* and professional standards require the credentialing body to provide examinees whose scores were invalidated the opportunity for a fair hearing of appeal or, at minimum, an opportunity to provide contrary evidence. In the current case, the test scores of OCPM examinees were invalidated as a group. Because the credentialing body and its testing contractor labeled the score invalidations a group action, they argued that the appeal language in the *Bulletin of Information* was not relevant. And, no testimony was presented that the OCPM examinees whose scores were canceled were given an opportunity to provide such evidence. Perhaps equally important is the fact that such evidence was solicited or considered.

The concept of *group invalidation* could well serve as the grist for an entire paper devoted to that topic. It is not the purpose of the this paper to delve deeply into that topic. A search for information related to group invalidation, however, suggests a conclusion that the concept is murky and one about which most relevant professional guidelines are silent or unclear. What is clear, as pertains to the current case, is that OCPM students registered for the examination as individuals, not as a group. Nothing in the *Bulletin of Information* provided to candidates suggested that any rights or responsibilities of candidates did not accrue to them as individuals. Further, invalidation of scores of the entire group has some, unknowable likelihood of invalidating the scores of those students (perhaps all) whose scores were obtained via legitimate preparation, making the opportunity for appeal of particular importance..

Thus, in the absence of language related to group invalidations in the *Bulletin of Information*, it appears that a responsibility existed on the part of NBPME/Chauncey to ascertain the validity of individual student's scores on an individual basis, and to adhere to the provisions of the contract with required an appeal hearing if properly requested.

Conclusions and Recommendations

This paper has examined four aspects of validation that arose in the context of a recent legal challenge to score invalidations involving a credentialing examination. Several lessons can be learned from the case regarding how credentialing organizations might protect themselves from successful legal challenges in the context of CAT/CBT (or many in paper-and-pencil contexts for that matter).

First, credentialing organizations must weigh the relative advantages and disadvantages of computerized test administration conditions as they relate to test security. If procedures such as a testing window--particularly in the presence of a shallow item pool--are used, organizations should consider monitoring disclosure of or inappropriate access to test content, either as an in-house function or as a contracted responsibility of the testing company or other vendor specifically focusing on test security (see, e.g., Caveon, 2004). Procedures for detecting and controlling item exposure in computerized administrations (see e.g., Lewis & McLeod, 1999; Lewis & Stocking, 2000) should be implemented.

Second, when conducting an investigation into suspicious test performance or when faced with the potential of invalidating scores, credentialing organizations must be cognizant of the responsibility to seek and consider information bearing on each potential interpretation. That is, both confirming and disconfirming evidence should, wherever practical, be sought and weighed.

Third, the process, data, and results of that weighing of confirming and disconfirming evidence--as well as process, data, and results of other related aspects of the score investigations should be well documented.

Finally, the language that appears in candidate informational materials can constitute an agreement between the organization and the examinee. Whatever provisions are contained in that agreement should be carefully described and--importantly--adhered to should disputes arise.

References

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

Caveon, Inc. (2004). Caveon test security. [Retrieved March 3, 2004 from www.caveon.com].

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it.* Mahwah, NJ: Lawrence Erlbaum.

Cizek, G. J. (2003). *Detecting and preventing classroom cheating*. Thousand Oaks, CA: Corwin.

Educational Testing Service. (2000). *Standards for quality and fairness*. Princeton, NJ: Author.

Ebel, R. L. (1961). Must all tests be valid? *American Psychologist, 16,* 640-647.

Gibley, Jr., C. (1998). The National Board of Podiatric Medical Examiners new testing methodology. *CLEAR Exam Review, 9*(2), 12-14.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535.

Ohio College of Podiatric Medicine (OCPM), et al., v. National Board of Podiatric Medical Examiners, (NBPME) et al., (2003). Case No. 1:02CV2435, U.S. Dist. Ct., Northern District of Ohio, Eastern Division.

Law.Com. (2004). Adhesion contract. Retrieved March 5, 2004 from http://dictionary.law.com/definition2.asp?selected=2325&bold=%7C%7C%7C%7C.

Lewis, C., & Sheehan, K. (1988). Computerized mastery testing. *Machine-Mediated Learning, 2*, 283-286.

Lewis, C., & McLeod, L. D. (1999). Detecting item memorization in the CAT environment. *Applied Psychological Measurement, 23*, 147-160.

Lewis, C., & Stocking, M. L. (2000). Methods of controlling exposure of items in CAT. In W. J. Van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 163-182). Dordrecht, Netherlands: Kluwer.

Lewis, C., & Wainer, H. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27*, 1-14.

National Board of Podiatric Medical Examiners [NBPME]. (2002). *Bulletin of information: 2002 examinations*. Princeton, NJ: The Chauncey Group International.

Table 1

Sample Format of Study Guide and Similar Item

Sample Study Guide Entries

---

5. TCA MOA?	block reuptake of NE, seratonin, and dopamine
18. What increases diffusion?	solubility
49. What happens if you remove lymphatics?	edema
67. Scurvy	Vitamin C deficiency
70. Flexes leg and thigh	sartorius

Item Identified as Matching Study Guide Entry #49

Removal of lower extremity lymph nodes in the treatment of malignancy increases the risk of which of the following:

A) tumor metastasis

B) chronic edema

C) keloid formation

D) fat embolism

Notes

1. Although a portion of the title of this paper ("Results...") might suggest that the legal challenge has been worked its way completely though the legal system, the decision in this case is currently under appeal.

2. In this location, and at various points throughout this paper, I made a conscious decision to ignore proper APA-style citation, quotation, or referencing technique. My purpose in writing this paper was to illustrate key validation concerns in licensure and certification contexts. I decided--as much as possible--to avoid attributing what might be considered to be oversights, inappropriate actions, etc., to specific persons or organizations, and to avoid attributions where they might be embarrassing or unnecessary. My assumption before and conclusion after involvement in this case is that the testing personnel and others involved were qualified, experienced, and deserving of professional respect. For the reader who desires verification of quoted material or other omitted information for scholarly purposes, I will provide full citations and references upon request.