1995

3. Policy Issues With Psychometric Implications

Michael Rosenfeld
Educational Testing Service, mrosenfeld@ets.org

Richard F. J. Tannenbaum
Educational Testing Service

Scott Wesley
Educational Testing Service

POLICY ISSUES WITH PSYCHOMETRIC IMPLICATIONS

Michael Rosenfeld

Educational Testing Service

Richard J. Tannenbaum

Educational Testing Service

Scott Wesley

Educational Testing Service

Testing candidates with disabilities, testing repeaters, and coaching involve issues of fairness, the validity of the inferences made from test scores, and protection of the public. Licensing boards must develop policies to deal with each of these issues. It is interesting to note that although all three are of concern to licensing agencies, little of the research on these topics has been conducted in licensure settings. This chapter discusses the results of research conducted on each topic, considers the psychometric implications for policy of each, and suggests steps licensing boards can take when formulating policy.

TESTING CANDIDATES WITH DISABILITIES IN LICENSURE SETTINGS

Disabled examinees take tests to apply for college, graduate school, and to be licensed or certified. Their ability to perform well on these examinations can be severely limited if the testing conditions or test format interact with their disability, but are not required for performance in school or on the job.

Most licensing agencies have been providing examinations in facilities accessible to disabled candidates, and have been providing alternative forms of examinations for many years (Schmitt, 1991). Accommodations for college-entrance examinations have been made since the 1930s (ETS, 1988). In 1937, a version of the Scholastic Aptitude Test was developed for students who are

visually impaired. The College Board, with the assistance of the American Foundation for the Blind, developed a braille booklet containing 100 antonyms, 50 analogies, and 50 reading comprehension items. A "talking book" record was also introduced which contained additional reading comprehension passages and questions. A braille practice booklet was developed to provide an opportunity for blind students to review the concepts covered by the test prior to taking the examination. Testing agencies had been providing accommodations to candidates from special populations, based primarily on the agencies' commitment to fairness and equal opportunity. The passage of the Americans with Disabilities Act (ADA) PL 101-336 now requires licensing agencies to provide appropriate accommodations for disabled test candidates. This legislation is likely to result in increased numbers of candidates requesting accommodations, and in licensing agencies providing them. The following section focuses on the requirements of the ADA that are related to testing, and the psychometric implications of these requirements.

The ADA

The ADA was enacted on July 26, 1990. It contains five major parts or titles. The act provides comprehensive civil rights protection to disabled individuals in the areas of employment, public accommodations, state and local government services, transportation, and telecommunications. Its intent is to increase job opportunities and access for disabled individuals. The testing requirements of the ADA took effect on January 26, 1992.

Title II of the ADA describes the responsibilities of state licensing agencies. It extends the prohibition of discrimination in federally assisted programs established by Section 504 of the Rehabilitation Act of 1973 (PL 93-112) to all activities of state and local governments, including those that do not receive Federal financial assistance. Title III delineates the responsibilities of private certification agencies. In general, the ADA emphasizes the need for (a) access to examination and course presentation facilities, (b) examination results that accurately reflect candidates' levels of knowledge or skill rather than their disabilities, and (c) administration of examinations for disabled candidates as often, and in as timely a manner, as examinations for nondisabled examinees. The section on examinations is quoted at length to provide examples of the language included in the ADA.

Section 36.309. This section delineates the ADA requirements for examinations and courses. It is part of Title III but also applies to state licensing agencies. The law reads:

> A. *General.* Any private entity that offers examinations or courses related to applications, licensing, certification, or credentialing for secondary or postsecondary education, professional, or trade purposes shall offer such examinations or courses in a place and manner accessible to persons with disabilities or offer alternative accessible arrangements for such individuals.
>
> B. *Examinations.*
>
> (1) Any private entity offering an examination covered by this section must assure that—

(i) The examination is selected and administered so as to best ensure that, when the examination is administered to an individual with a disability that impairs sensory, manual, or speaking skills, the examination results accurately reflect the individual's aptitude or achievement level or whatever other factor the examination purports to measure, rather than reflecting the individual's impaired sensory, manual, or speaking skills (except where those skills are the factors that the examination purports to measure);

(ii) An examination that is designed for individuals with impaired sensory, manual, or speaking skills is offered at equally convenient locations, as often, and in as timely a manner as are other examinations; and

(iii) The examination is administered in facilities that are accessible to individuals with disabilities or alternative accessible arrangements are made.

(2) Required modifications to an examination may include changes in the length of time permitted for completion of the examination and adaptation of the manner in which the examination is given.

(3) A private entity offering an examination covered by this section shall provide appropriate auxiliary aids for persons with impaired sensory, manual, or speaking skills, unless that private entity can demonstrate that offering a particular auxiliary aid would fundamentally alter the measurement of the skills or knowledge the examination is intended to test or would result in an undue burden. Auxiliary aids and services required by this section may include taped examinations, interpreters or other effective methods of making orally delivered materials available to individuals with hearing impairments, brailled or large print examinations and answer sheets or qualified readers for individuals with visual impairments or learning disabilities, transcribers for individuals with manual impairments, and other similar services and actions.

(4) Alternative accessible arrangements may include, for example, provision of an examination at an individual's home with a proctor if accessible facilities or equipment are unavailable. Alternative arrangements must provide comparable conditions to those provided for nondisabled individuals. (pp. III-100-103)

Definitions of disability. Section 36.104 contains the ADA definition of disability. This is quite broad, and describes which individuals are covered under the ADA. The law reads:

Disability means, with respect to an individual, a physical or mental impairment that substantially limits one or more of the major life activities of such individual; a record of such an impairment; or being regarded as having such an impairment.

(1) The phrase *physical or mental impairment* means—
   (i) Any physiological disorder or condition, cosmetic disfigurement, or anatomical loss affecting one or more of the following body systems: neurological; musculoskeletal; special sense organs; respiratory, including speech organs; cardiovascular; reproductive; digestive, genitourinary; hemic and lymphatic; skin; and endocrine;
   (ii) Any mental or psychological disorder such as mental retardation, organic brain syndrome, emotional or mental illness, and specific learning disabilities;
   (iii) The phrase physical or mental impairment includes, but is not limited to, such contagious and noncontagious diseases and conditions as orthopedic, visual, speech, and hearing impairments, cerebral palsy, epilepsy, muscular dystrophy, multiple sclerosis, cancer, heart disease, diabetes, mental retardation, emotional illness, specific learning disabilities, HIV disease (whether symptomatic or asymptomatic), tuberculosis, drug addiction, and alcoholism.
   (iv) The phrase *physical or mental impairment* does not include homosexuality or bisexuality.

(2) The phrase *major life activities* means functions such as caring for one's self, performing manual tasks, walking, seeing, hearing, speaking, breathing, learning and working.

(3) The phrase *has a record of such an impairment* means has a history of, or has been misclassified as having, a mental or physical impairment that substantially limits one or more major life activities.

(4) The phrase *is regarded as having an impairment* means—
   (i) Has a physical or mental impairment that does not substantially limit major life activities but that is treated by a private entity as constituting such a limitation;
   (ii) Has a physical or mental impairment that substantially limits major life activities only as a result of the attitudes of others toward such an impairment; or
   (iii) Has none of the impairments defined in paragraph (1) of this definition but is treated by a private entity as having such an impairment.

(5) The term *disability* does not include—
   (i) Transvestism, transsexualism, pedophilia, exhibitionism, voyeurism, gender identity disorders not resulting from physical impairments, or other sexual behavior disorders;

(ii) Compulsive gambling, kleptomania, or pyromania; or
(iii) Psychoactive substance use disorders resulting from current illegal use of drugs. (Equal Employment Opportunity Commission and U.S. Department of Justice 1991, pp. II-16-20)

Discussion of board responsibilities. As can be seen, the ADA describes disabilities quite broadly. It also describes two general types of accommodations. The first involves the accessibility of facilities to individuals (e.g., wheelchair accessibility); the second involves modifications to the examination itself or the examination process (e.g., providing additional time to take the examination or using of large-size print). The ADA requires that decisions concerning accommodations be tailored to the individual needs of the candidate and the essential functions of the job. The decision made by the licensing or certification board should be designed to provide the candidate an opportunity to demonstrate his or her knowledge and skill on as equivalent a basis as possible. (In many instances, the request for a particular accommodation will initially be made by the candidate and then verified by an appropriately licensed professional or a certified specialist selected by the candidate.)

A board must make several types of decisions when considering an applicant with a disability. First, the candidate must have the same qualifications to take the examination as all other candidates. Examples of such qualifications include educational attainment and work experience. This is consistent with the ADA's concept of a qualified individual with a disability (p. II-26). The Act clearly states that a person must be qualified to perform the job in question, with or without a reasonable accommodation. Second, the board must decide if the disability will affect the candidate's ability to perform the essential functions of the job. For example, it would be unreasonable to expect a candidate who cannot see to perform surgery or function as a building inspector because both jobs are heavily dependent on visual ability. Once the board has decided a candidate is qualified to take the examination and can perform the essential functions of the job, it must determine what modifications in the examination or the examination process it is willing to make to allow the candidate a fair opportunity to demonstrate relevant knowledge or skills.

ADA regulations provide two criteria licensing and certification boards can use in making decisions about accommodations for disabled candidates. The first would require the board to determine whether it believed the accommodation would fundamentally alter the measurement of the construct being assessed. For example, if a test were designed to measure reading comprehension and the accommodation requested was to allow someone to read the test aloud to the candidate, the accommodated test would measure listening comprehension, not reading comprehension. The inferences made about the test score would thus be invalid. The second criterion involves whether the board believes the accommodation represents an "undue burden" because of the cost or difficulty in developing or administering the modified examination. Clearly, applying the ADA to individual situations requires sound professional judgment.

Types of Accommodations. Paragraph 36.104 of the ADA delineates the types of physical and mental disabilities covered by the Act. These definitions are, for

the most part, taken from Section 504 of the Rehabilitation Act of 1973. Many licensing, certification, and admission-testing agencies already provide accommodations to candidates who are physically disabled, blind or visually impaired, deaf or hard of hearing, learning disabled, or mentally disabled. In many of these categories the nature and severity of the disability varies greatly from candidate to candidate. Therefore, no single accommodation is likely to be appropriate for all members of any group of disabled candidates. Listed below are some testing accommodations that are commonly made available to disabled test candidates.

Alternative Test Versions. Many tests can be provided in braille, large print, and audiocassette versions. Sometimes test questions in the print version may have to be reformatted, substituted, or dropped from the examination because they are not appropriate for the specific disability (e.g., a visual stimulus or test question that cannot be translated into braille). Alternative ways to record answers to test questions have also been provided. These include allowing the use of typewriters or computers rather than the typical machine-scorable answer sheets. Answers can be written on the test booklet itself and on large-print answer sheets.

Assisting Personnel. When special versions of a test are not available, it is not uncommon for testing agencies to provide or allow for candidates with disabilities to use a reader. Amanuenses may be used by disabled candidates to help them record their answers. Deaf or hard-of-hearing candidates whose primary mode of communication is sign language may need an interpreter.

Assisting Devices. Some assisting devices can be used. These might include an Opticon, Visualtek, or a braille typewriter for a print test, or a voice synthesizer or a special keyboard for a computer-based test.

Separate Testing Locations. Tests that are usually group administered have frequently been provided to disabled individuals in a separate room or at a separate site. This is particularly true if extra time is needed, a reader or amanuensis is used, or if the test is in braille or on a cassette. A separate room could also provide a disabled examinee an opportunity for more space, the use of enhanced lighting, special seating, and provisions for rest periods.

Extra Time. Most standardized tests are administered so all candidates have the same amount of time to respond to the test questions. Some accommodations provided to disabled candidates, such as the use of a cassette or braille version of the test, or the use of a reader, may require more testing time. In addition, some individuals with physical or mental disabilities may require time to rest during the examination or between sections of the examination. Extra time is the accommodation most frequently provided in licensing as well as other testing contexts.

Appropriate and Inappropriate Accommodations. Accommodations provide an accessible alternative way for the disabled candidate to demonstrate the desired skill. Accommodations are intended to provide an equally accurate assessment of the knowledge, skill, or ability that the test is designed to measure for both disabled and nondisabled candidates. For example, a candidate with a visual disability may take a reading comprehension test in braille or using large print, and the test would still measure reading comprehension. This accommodation provides a format change that allows the disabled candidate to demonstrate the desired ability

unimpaired by the candidate's disability. This would be considered an appropriate or, as Phillips (1993) refers to it, a valid accommodation. The inference made concerning reading ability would be similar for candidates taking the braille version of the test and those taking the test in its standard print version.

An inappropriate or invalid accommodation is one in which the accommodation changes the construct being measured. As in a previously mentioned example, if the purpose of a test was to assess a candidate's reading comprehension, and the candidate requested that the test be read to him or her, the accommodated test would measure listening comprehension, not reading comprehension.

Boards should exercise care when deciding which accommodations to offer or allow. They must keep clearly in mind the purpose of the test, what it is designed to measure, and the inferences that are to be made from the test scores. Before making a final decision, the board might do well to consult with psychometric and legal professionals.

Many accommodations can be provided that will not affect the underlying construct being measured. Boards have the right to deny requests they believe could alter the construct, however. Licensing boards have the dual responsibility to provide reasonable and appropriate accommodations to disabled examinees while providing protection for the health, safety, and welfare of the general population.

Psychometric Implications of Test Accommodations. Accommodations for disabled candidates called for in the Rehabilitation Act of 1973 and the ADA reflect the first instances in which testing organizations have been required to modify testing conditions or the format of an examination for a particular subgroup of test takers. This raises a number of measurement issues. For example, can the scores obtained from an accommodated and a standard administration be equated? Do the scores have the same meaning as in a standard administration? Should the scores obtained from an accommodated test administration be noted or "flagged" so those responsible for using test scores are aware that an accommodation has been provided to a disabled candidate? These concerns are discussed below.

Equating Scores. Can the scores obtained from a test administered with special accommodations be equated with those from a standard test administration? This issue is discussed in "The Score" (APA, 1993), the newsletter of the Division of Evaluation, Measurement, and Statistics of the American Psychological Association. It discusses various equating strategies and the technical difficulties associated with each approach.

One major problem is that the two groups being compared are not random samples from the same population. Secondly, the two groups are not as nearly equivalent as could be desired; the disability may have affected the educational experience and learning of one of the groups. Thirdly, the testing conditions differ: The accommodation may have provided more time, or a different item format. Under these "new" conditions, the construct being measured may have changed even though the nature of the change may not be as obvious as the example noted earlier of shifting from the measurement of reading comprehension to the measurement of listening comprehension.

These problems make it very difficult to equate the scores of examinees taking a test under standard conditions with those of examinees taking the same test with special accommodations. "The Score" concludes, "There is no standard technical solution available for precisely equating a modified administration of a cognitive test, which has itself been modified, to the standardized form—at least, in those situations where the modification is one that will have an effect on test scores" (APA, 1993, p. 8).

Meaning of Scores. The second issue is whether scores on a modified test have the same meaning in terms of what they measure and how they measure it. Standard 14.6 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1985) states that "When feasible, the validity and reliability of tests administered to handicapped people, with and without accommodation, should be investigated" (p. 80). However, such studies have rarely, if ever, been conducted in the areas of licensing and certification. There are usually too few candidates requesting accommodations in any one program to make it feasible to conduct studies of this sort within a short time span. Often, it takes the accumulation of data over many years to answer questions of this type. Data are available, however, from the area of college admissions testing. A report from a National Academy of Sciences Panel (Sherman & Robinson, 1982) called for research to clarify whether tests modified for examinees with disabilities are comparable to standard tests, and whether they give valid estimates of the academic abilities of disabled people.

A series of studies on the Scholastic Aptitude Test (SAT) and the Graduate Record Exam (GRE) General Test were undertaken jointly by the College Board, Educational Testing Service, and the Graduate Record Examination Board in response to the National Academy of Science Panel report (Willingham, Ragosta, Bennett, Braun, Rock, & Powers, 1988). The studies cover four major groups of people with disabilities (deaf and hard of hearing, learning disabled, physically disabled, and visually impaired students). Several indicators of score comparability were discussed. Those judged relevant for licensing and certification are summarized below:

The internal consistency reliability of individual subscores for the standard SAT and GRE tends to be approximately .90. The reliability of these tests when administered with accommodations to disabled students was approximately the same. The standard error of measurement was virtually the same for the disabled groups and for those taking the tests under standard conditions.

The factor structure of the SAT and GRE were very similar for several different groups of disabled and nondisabled examinees. This result indicates that nonstandard tests (tests with accommodations) have comparable meaning for the cognitive abilities they measure.

There was little evidence of differential item difficulty. It appears the SAT and GRE are largely free of item types that are unusually difficult for students with particular disabilities compared with other items measuring the same ability.

The use of test scores was studied as another aspect of comparability, namely, admission decisions of colleges and universities using the SAT. Although admissions decisions are not directly relevant to licensing, the use of flagged test scores should interest licensing boards. Willingham et al. concluded that the nature of the

selection process seemed comparable for nondisabled and disabled applicants submitting flagged scores, based on an analysis of decisions using test scores and school grades. The probability of admission increased for both groups of applicants as test scores and grades increased. The weight placed on these measures seemed similar for both groups.

When academic performance was predicted using both test score and prior grades, there was little consistent over- or underprediction for the four categories of disabled students. However, the academic performance of some categories of disabled students was less predictable than that of nondisabled students from test scores, from previous grade-point averages, or from both combined. The performance of three of the four groups of disabled students was significantly under- or overpredicted when predictions were based on test scores alone. Deaf and hard-of-hearing students were underpredicted by the SAT; physically disabled and learning-disabled students were overpredicted.

There was evidence that nonstandard timing versions of the SAT and GRE were not comparable to the standard version. All groups of disabled candidates were more likely to complete the test. Some items near the end of the test were easier for three of the four disabled groups studied; and some instances of overpredicted college performance suggested that extended testing time may have contributed to inflated test scores.

Another study (Laing & Farmer, 1984) conducted by the American College Testing Program (ACT), investigated the equivalency of examination formats for examinees with disabilities (physical, learning, visual, and auditory) and nondisabled examinees using standard examination formats. Data from high school students taking the ACT assessment for college admission were used in the study. ACT identified 880,040 examinees who were tested on national test dates in 1982–83, of which 1% (6,289) indicated they had a disabling condition that might require related services. Visually impaired examinees obtained the highest test scores, and deaf and hard-of-hearing examinees obtained the lowest test scores of the disabled groups. These findings are consistent with those from other studies (Bennett, Ragosta, & Stricker, 1984; Ragosta & Kaplan, 1986) which found that visually impaired students and physically disabled students obtained higher mean SAT scores than did learning disabled students, who obtained higher mean scores than deaf and hard-of-hearing students. Scores for disabled examinees in the ACT study, even with accommodations, were lower than those received by nondisabled examinees. This was true for all groups except for visually impaired examinees given accommodations during testing. The prediction of grades was generally lower for disabled examinees. However, caution was recommended in interpreting the results, given small sample sizes and the reliability of self-reported high school and college grades.

The results provided above indicate that nonstandard versions of the SAT and GRE were comparable to standard versions with respect to reliability, factor structure, and item functioning. For the SAT, the use of test scores and grades for admissions decisions was also comparable. (Because of limited sample size, a similar study could not be conducted using GRE scores.) Although there seemed

little systematic over- or underprediction of academic performance when both SAT score and previous grades were used, there were instances of over- and underprediction for three of the four disabled groups when test scores were used alone. There was also evidence that nonstandard timing versions of the SAT and GRE were not comparable to the standard version. Although the results from admissions testing provide some indications of comparability, the findings are not definitive.

What are the implications of the research for licensing boards? The results cited above were obtained within an admissions-testing context by organizations that have some of the largest examinee populations in the world. Even these organizations had difficulty conducting some aspects of their studies because of limited sample size and problems with criterion measures. The results presented are based on the best data currently available to investigate the comparability of test scores of disabled candidates taking examinations under nonstandard conditions with nondisabled candidates under standard conditions. It should be noted that these studies were conducted with multiple-choice items and were predominantly measures of verbal and quantitative abilities. There were no results presented on performance assessment, computer-based assessment, or constructed-response measures. In terms of their usefulness for the licensing context, these studies can only be considered suggestive. Comparability studies will be extremely difficult for licensing boards to conduct, however, given the relatively small number of candidates tested overall and the still smaller number who are tested with particular types of disabilities and different accommodations. We do not have definitive answers now about the comparability of test scores obtained under standard and nonstandard conditions for these two groups of examinees, and we are not likely to have them in the near future. It is important that licensing boards collect data in order to accumulate enough information over time to conduct research studies on this issue.

Flagging Test Scores. Because we do not know whether scores obtained for disabled examinees in a licensing context are directly comparable to the scores obtained by nondisabled examinees under standard conditions, should the scores obtained by disabled examinees under nonstandard conditions be flagged? Standard 14.2 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1985) states that "until tests have been validated for people who have specific handicapping conditions, test publishers should issue cautionary statements in manuals and elsewhere regarding confidence in interpretations based on such test scores" (p. 79). This is stated as a primary standard. Although the ADA does not prohibit the practice, many candidates with disabilities perceive flagging as discriminatory. It seems that licensing boards may have a responsibility to flag test scores until validity studies have been conducted. The questions licensing boards must answer include:

- Should test scores be flagged?
- If so, under what conditions?
- Who should have access to this information?

The purpose of flagging a test score is to inform and caution users that the score was obtained under nonstandard conditions and might not have the same meaning

as other scores obtained under standard conditions. The board should consider who uses the test score other than the board itself, and whether the flag would prevent an inappropriate decision being made with that score.

One rationale for flagging a test score would be research purposes, because it is clear that more research must be conducted on the comparability of test scores taken under standard and nonstandard conditions. As numbers of candidates with various types of disabilities accrue, it is important for licensing boards to investigate the comparability of scores. The possibility of future litigation presents another reason for boards to keep records of the number of disabled examinees who have received accommodations and the type of accommodations provided. Flagged scores could be kept secure at the licensing board and used only for research and record keeping.

Because one of the major responsibilities of licensing boards is to protect the public from practitioners who lack the minimum qualifications for competent performance (Shimberg, 1985), boards should consider if flagging would help protect the public. In this regard, a board has responsibility for deciding who is eligible to take its licensing examination (Shimberg, 1985). If applicants requesting a particular accommodation are required to specify the nature of their disabilities, the board must decide whether candidates will be able to perform the essential functions of a given job, and whether the proposed accommodation would fundamentally alter the construct being measured. This action would be consistent with the content validity model used to support most licensing examinations (Impara & Stoker, 1985; Kane, 1982; Shimberg, 1981). If the board believes the nature and extent of the disability will not allow the examinee to perform essential functions of the job, or that the accommodation will alter the construct being measured, it is the board's responsibility to inform the examinee that he or she is ineligible to take the licensing examination. If the board decides the candidate is eligible to take the examination and the accommodation is acceptable, the board has agreed that this is an appropriate way for the examinee to demonstrate possession of the knowledge and skill necessary to perform the essential functions of the job for which the license is being issued. Under these conditions, it would seem there is little or no basis for flagging the score other than for the board's own records as described above.

Boards should carefully decide whether they believe it necessary to flag scores and, if so, to document the rationale for their decision. If scores are flagged, the board should develop policies and procedures designed to protect the rights of disabled candidates and insure that the scores are kept secure from unauthorized personnel and uses. Flagged scores should not be used in a way that discourages eligible candidates from requesting accommodations, nor that harms their opportunity for employment.

Summary and Implications for Licensing Boards

On July 26, 1992 the ADA went into effect, requiring licensing boards to modify testing conditions and/or formats for disabled individuals requesting accommodations. This was clearly a social policy decision, but it does raise a number of psychometric issues regarding how to implement this policy while maintaining

standards and test score comparability. Unfortunately, the quality and quantity of research data on disabled examinees are very limited. As a result, the possibility of establishing the comparability of nonstandard test scores on the basis of empirical studies alone is also limited. It appears that licensing boards will need to use logical analysis and sound judgment to decide what constitutes a comparable task for a disabled examinee, taking into account the purpose of the test as well as the degree of the disabling condition.

Standardization was developed to increase the likelihood that all examinees would have an equal opportunity to demonstrate the relevant knowledge and skills and to provide a common basis for interpreting test scores. Thus, the purpose of standardizing the testing task was to make it more objective and fair for all candidates. If for some examinees, however, the task has extraneous sources of difficulty because of their disability, the test would be unfair. The goal of the accommodation, then, is to eliminate or greatly reduce the extraneous sources of difficulty. One can consider a special accommodation as an attempt to modify the test or the testing condition so it provides comparable information about the individual on the construct the test is designed to assess. In the absence of a great deal of empirical data, this will require the exercise of sound professional judgment.

Boards must balance their responsibility to provide access and accommodations to disabled examinees with their responsibility to protect the health, safety, and welfare of the general population.

In addition, the board must decide, for each disabled examinee requesting an accommodation, whether the:

- Candidate has met all qualifications to take the examination.
- Disability will affect the candidate's ability to perform essential functions of the job.
- Accommodation would alter the measurement of the construct being assessed.
- Accommodation is available and feasible without placing an undue burden on the board.

Boards must make good-faith efforts to meet both sets of demands and, as case law evolves under the ADA, must track rulings and modify their policies and procedures accordingly.

Table 1 presents some steps boards can follow to assist in making these decisions.

TESTING REPEATERS

It is probably safe to say that not all candidates who take a licensure test will pass. Some candidates may not pass the test because they lack the requisite knowledge or skills being measured by the test. Others may not pass because of chance factors unrelated to the purpose of the test (e.g., high test anxiety, temporary illness, or fatigue). Although the reasons for candidates not passing may be varied (and, no doubt, readers have thought of many more than we listed), one thing all such candidates have in common is the need to repeat, that is, to take the licensure test again (provided, of course, that they still want to enter the particular profession).

Table 1. Suggestions for Setting Policy on Disability Issues

1.	Prepare up-to-date job analysis information that can be used to establish the essential functions of the particular job or profession in question.
2.	Develop and publish a policy on examination accommodations with the advice of psychometricians and legal counsel.
3.	Decide on the written documentation necessary to request an accommodation. It would be wise to request an adequate description of the disability, evidence that the disability currently exists, and a rationale for the accommodation requested. This documentation should be provided by an appropriate licensed professional or certified specialist.
4.	Establish procedures for responding to requests for accommodations in a timely manner.
5.	Identify consultants expert in various disabilities to assist in reviewing and assessing documentation and to perform applicant evaluations when necessary.
6.	Develop procedures for board review of all requests for accommodations, or at least those requests which are denied.
7.	Keep a record of all requests for accommodations and the response to each request.
8.	Decide whether to flag scores, and document the rationale for the decision.
9.	Track the emerging court cases under the ADA to determine whether board policies and procedures are consistent with case law.
10.	Produce additional program materials and procedures needed to develop special test editions, to administer tests, and to provide services for disabled examinees. Steps should also be taken to develop practice test materials for disabled examinees.
11.	Maintain records for possible use in research activities or litigation.

Simply letting those who do not pass take the licensure test again—after all, we all deserve at least a second chance—like most life events is not without complications. In this section, we focus on one potential measurement confound associated with testing repeaters: the practice effect.

Practice Effects and Validity Implications. A practice effect is defined as a gain in test performance resulting from previous experience with the same test or a parallel (alternate) form of the test (Weiss, 1961). Unlike coaching (discussed in a later section of this chapter), in which candidates participate in test preparation activities specifically to improve their test scores, the benefit from practice

is derived solely from familiarity with the test and the testing situation. (Candidates who have repeated and/or who have been coached have a greater advantage than first-time test takers who have not been coached. To reduce this advantage as well as to promote test fairness, many testing organizations provide all candidates with a pre-examination booklet that includes sample test items and general test-taking strategies.)

As with all testing applications, at issue here is validity, or accuracy of the inferences drawn from the scores obtained on the licensure test. Licensure tests are designed to ensure that candidates who seek to enter a profession possess knowledge and skills necessary to protect the public's health, safety, and welfare (*Standards for Educational and Psychological Testing*, AERA, APA, & NCME, 1985). The objective is to determine whether candidates have minimal competence; licensure testing, as such, is a selecting-out process (Madaus & Mehrens, 1990). In the vernacular of decision theory (Cronbach & Gleser, 1965), licensure testing also attempts to minimize the incidence of both false acceptances and false rejections; that is, to reduce the granting of licenses to those who lack minimal competence and to avoid withholding licenses from those who possess minimal competence.

The validity of test scores will be compromised to the extent that practice effects are large. A gain in a test score, due only to the effects of practice, would incorrectly be attributed to increased knowledge or improved skills. The social consequence of this false inference takes on much greater import if the spurious gain results in a test score that exceeds the cut score established for the licensure test. The explicit intention of licensure testing would be circumvented if a professional license was granted to a candidate who did not possess the knowledge and skills necessary to safeguard the welfare of the public. It is critical, therefore, that the effects of practice on licensure testing and the factors that contribute to and moderate these effects be better understood. To this end, we will attempt to delineate the domain of practice effects as it relates to licensure testing, bearing in mind that in doing so, we may raise more issues than answers.

Practice Effects: A Brief Review. Researchers investigated the effects of practice on intelligence tests as early as the 1920s (e.g., Dunlap & Snyder, 1920; Richardson & Robinson, 1921; Thorndike, 1922). Though the explanations for the obtained results were not always consistent, the general finding was. Test scores increased upon retesting.

One of the first reviews of literature on the effects of practice was carried out by Weiss, who reviewed 17 studies conducted in Great Britain and the United States on tests of mental ability and scholastic aptitude (1961). He concluded that: (a) practice improved performance; (b) significant practice effects occurred on a first and second retest, but the effects diminished after that; (c) practice effects varied with the time between test administrations—significant effects were obtained for time intervals of 2 weeks to 3 months; and (d) practice effects interacted with mental ability—more intelligent test takers appeared to benefit most from practice.

Since the time of the Weiss review, other studies have attempted to explicate more fully the domain of practice effects. Attention began to focus on character-

istics of the test and the testing process that practice affected. As was the case with previous studies, however, the preponderance of tests included in these studies were either mental aptitude or achievement tests. None were used for professional licensure. And most, if not all, used a traditional multiple-choice item format.

Rock and Werts (1980) examined the effects of practice on the Graduate Record Examinations (GRE) Aptitude Test. They were particularly interested in the effects of time and gender on repeaters' performance. They found, in general, that test scores on both the verbal and quantitative components increased upon retesting, regardless of the gender of the test taker. Slightly greater gains after one retest were observed on the verbal component (about 26–27 points) compared with the quantitative component (about 23 points). Both men and women single-repeaters showed greater gains in their verbal scores as the length of time between test administrations increased. This was attributed to growth in verbal abilities over time, not just to the effects of practice. The same result was not observed, however, for the quantitative component. As noted by Rock and Werts, verbal skills would appear to increase throughout adulthood, whereas quantitative skills would appear to be relatively stable.

Wing (1980) examined the effects of practice on five abilities (verbal, judgment, induction, deduction, and number) as measured by the Professional and Administrative Career Examination (PACE), a test used by the federal government to select entry-level employees. Data were collected from more than 60,000 test takers. The effects of practice were found to vary depending upon the ability being measured, the order of presentation of the items, the difficulty of the items, and the speededness of the items.

Wing concluded that practice effects were (a) largest for item types (e.g., letter series, geometric classifications, arithmetic reasoning) that were solvable by systematic application of general problem-solving skills; (b) next largest for test parts subject to speededness; and (c) smallest for item types (vocabulary, comprehension) solvable by application of previously acquired general information.

In 1984, Kulik, Kulik, and Bangert conducted a meta-analysis of 40 studies to identify variables that had an impact upon practice effects. Among the variables of interest were the ability level of the subjects (high, medium, or low); the grade level of the subjects (elementary, high school, postsecondary); and the type of test used (aptitude versus achievement).

Their analyses revealed that practice effects (as measured by an effect-size statistic) were larger when the tests were identical than when the tests were parallel forms of one another (though the effect was still significant in the latter case). The effects of practice were also positively related to the number of practice tests. The average effect size increased from .42 from one practice on an identical test to 1.89 for seven practice tests. For parallel forms, the average effect size increased from .23 to .74. Lastly, the magnitude of practice effects was related to the ability level of the test takers. High-ability test takers gained more from a single practice test (effect size = .82) than did middle-ability test takers (effect size = .40) and low-ability test takers (effect size = .17). Neither grade level nor type of test significantly affected the magnitude of practice effects.

The most recent synthesis of the literature on within-test practice effects for aptitude tests was conducted by Powers (1986). Within-test practice refers to previous exposure to item types that appear later in the same test. Powers coded studies according to the seven characteristics of test items: (a) number of response options, (b) option format, (c) item difficulty, (d) time per item, (e) length of test directions, (f) examples, and (g) overall complexity of directions and/or task. He then related practice effects (as measured by an effect-size statistic) to the item characteristics.

Practice effects were found to be highly related to both the length of directions ($r = .49$) and the complexity of directions ($r = .63$). Likewise, practice effects were related to option format ($r = .42$). In particular, fixed-format items (those in which the same set of alternative answers was used for each question) were associated with the larger effects. In addition, significant relationships were obtained between the number of response options and practice effects ($r = .40$) and between the time allotment per item and practice effects ($r = -.40$). In the latter case, the greater time per item was associated with smaller practice effects (cf. Wing, 1980).

Perhaps the only study to examine the effects of special test preparation on constructed-response items was conducted by Powers, Fowles, and Farnum (1993). Though actually a study of coaching effects, its results are noteworthy, and may be viewed as an upper limit of the effects of practice alone. A pool of 10 essay topics was disclosed and used for coaching purposes by instructors at four different colleges or universities. Following the coaching, students wrote two essays—one on a previously disclosed topic and the other on a topic that was not included in the disclosed set. Scoring of the essays was done by trained readers who independently assigned holistic scores on a 6-point scale. The results indicated relatively small differences between the scores on the disclosed essay and the new essay topics (across all students, the effect size was .15). Furthermore, using a cut score of 3.0, Powers et al. found little increase in the pass rate as a result of students writing on a disclosed topic compared to a new topic.

Summary

Several generalities may be culled from research on the effects of practice (also see Bond, 1989; Hopkins, Stanley, & Hopkins, 1990):

- Practice effects are greater on identical forms of a test than on parallel forms of a test.
- The average practice effect for a group of test takers is approximately .20 standard deviation units.
- Test takers of high ability benefit most from practice.
- Practice effects are more pronounced on speeded tests than they are on power tests.
- Less-experienced test takers benefit most from practice.
- The longer the time interval between the test and the first retest, the smaller the effects of practice (exclusive of growth effects).
- The more complex the item, the greater the effects of practice.

- Certain types of items (e.g., constructed-response) may be more resistant to practice effects than traditional multiple-choice items.

Practice Effects and Licensure Testing

Tests of professional licensure are noticeably missing from the research on practice effects. We can only speculate this may be because of the smaller numbers of test takers compared, for example, to Scholastic Assessment Test takers; or because the failure rate in licensure testing may not be high enough to prompt the concern of licensing agencies.

We would rather err on the side on conservatism and assume that licensure tests are prone to the effects of practice, at least to some degree. The interpretation of the significance of these effects, however, may need to be viewed differently for licensure tests. Unlike most aptitude or achievement tests, licensure tests are criterion referenced. That is, test scores are compared to an external cut score; test takers' scores are not compared to one another. The real issue, then, is not whether there is a practice effect per se, but whether the effect is strong enough, on average, to push the test taker above the cut score on repeated administrations of the licensure test or alternate forms thereof. This question awaits empirical investigation.

Psychometrically Based Issues Related to Testing Repeaters

Conjoined with the issue just raised are a variety of psychometrically based concerns. In this section we will acquaint the reader with some of these concerns. (Where appropriate, the reader will be directed to other chapters in this book for more in-depth discussions of these psychometric issues.)

Cut Scores (also see chapter 10). A cut score or passing score is typically set by a committee of subject-matter experts using any of a number of standard-setting procedures (e.g., Angoff, Jaeger, Nedelsky, contrasting groups). In order to diminish the effects of practice, emphasis must be placed on setting a cut score that unambiguously differentiates between those candidates who do and do not possess minimal competence. Measurement error should be explicitly considered during the standard-setting process. The standard error of the cut score should be such that the rates of false rejections and false acceptances are minimized.

Regression Effects. It is probable that upon retesting, a candidate's test score will increase, due, in part, to simple regression effects (Campbell & Stanley, 1966). That is, candidates who have scored very low on the initial test will, on average, score higher upon retesting (i.e., their scores will regress towards the mean score of the second test). This phenomenon occurs because of the imperfect correlation between the two tests. Without recognizing the potential impact of regression effects, the inference drawn from a test score above the cut score—that a candidate possesses minimal competence—may be suspect.

Equating (also see chapter 11). Testing repeaters may also affect both the methods used for equating and the outcomes of equating studies. Essentially, equating refers to statistical procedures designed to ensure that scores from alternate forms of a test will be directly comparable (Angoff, 1971). A frequently used equating design for licensure testing is the nonequivalent groups-common

item method. In this design, an identical subset of test items appears in each form of a test along with a distinct subset of test items. Two groups of test takers receive each form of the test. The comparability of the test scores is based upon the results obtained for the common (equated) subset of test items. If a large proportion of repeaters were included in the equating study, however, their previous exposure to the equated subset of test items would introduce an unwanted source of error.

The presence of a large number of repeaters in the second test administration would most likely lead to a gain in scores on the equated subset of test items. This could lead to the erroneous conclusion that the test takers in this administration have higher abilities than the group in the previous administration. A related confound arises if the nonequated items in the second test administration now appear to be more difficult than the nonequated items in the first test administration. A likely, though erroneous, outcome would be that the cut score for the second test administration is adjusted downward to compensate for the perceived greater difficulty of the items that constitute the second test.

Another form of equating, section pre-equating (Holland, 1981) does not require the use of two complete forms of a test; rather, multiple sections of items for equating are embedded across operational tests. Not all candidates, therefore, receive the same equating sections. The placement of the equating sections also varies across the operational tests; and the equating sections do not count toward the candidate's test score. Though promising, this method of equating may be prone to within-test practice effects. That is, because each pre-equating section is parallel to some operational section of the test, candidates may receive practice on particular item types that will affect their performance on the scored sections. The magnitude of these effects may vary depending upon the types of items (see Leary & Dorans, 1985, for a review of within-test effects).

Test Security. According to Burns (1985), for licensure testing to be considered secure, all candidates should have the same testing experience, and some candidates should not gain advantage by prior knowledge of the test. Repeaters clearly gain advantage by their prior exposure to and experience with either the same test or an alternative form of the test, however. And, as Burns notes, licensure tests may be particularly vulnerable to breaches of security because their specialized content may not readily lend itself to the construction of large item pools. It would appear, then, that part of maintaining the security of licensure testing is reducing the effects of previous exposure to the test (i.e., practice effects).

Time between test administrations. One of the easiest ways to reduce the effects of practice and to enhance test security is for the licensing agency to set a minimum interval before a candidate is eligible to repeat. Candidates may be required to wait a minimum of 6 months before being allowed to repeat, for example. Safeguards, such as verifying candidates' identities, could be implemented to ensure that candidates are not taking the licensure test before they are officially permitted to do so.

Item types. As we have seen, research has indicated that practice does not affect all item types similarly. Items that are not speeded are less prone to practice effects, for example, as are items not solvable by the application of specific rules.

Less complex items also appear more resistant to the effects of practice. Using constructed-response types of items may reduce the effects of practice. Continued efforts are needed to clarify the characteristics of items that make them resistant to the effects of practice.

Alternate forms. The effects of practice may be reduced, (though as noted earlier, not eliminated) by using multiple forms of the licensure test. Practice effects are less pronounced when alternate forms of a test are used. One effective variant of alternative forms testing is called spiralling. This refers to the packaging and subsequent distribution of multiple forms of a test to an administration site. By spiralling the tests, essentially random groups of test takers receive an alternate form of the test. The chances of a repeater receiving the same form more than once are thus dramatically reduced.

Computerized adaptive testing. Computerized adaptive testing (CAT) is a fairly recent technological development that may prove useful to reduce the effects of practice and increase test security. Adaptive testing was designed to enable more accurate and more efficient determinations of a test taker's true ability by matching the difficulty level of each presented item to the estimated true ability level of the test taker (Lord, 1980).

In CAT, as described by Wainer (1990), a test taker begins the test with an item in the middle of a prospective range of difficulty. Then, depending upon the correctness of the response, the next item is either harder or easier. If the item was answered correctly, the next item would be harder; if, however, the item was answered incorrectly, the next item would be easier. After each response to an item, the test taker's current ability level is estimated. Based upon the current ability estimate, a new test item of appropriate difficulty is then selected. Testing continues in this manner until a predetermined level of measurement precision is attained, a preselected number of items has been given, or a predetermined amount of time has elapsed (Thissen & Mislevy, 1990). The most recent estimate of a test taker's ability level is used as the test score.

A particularly appealing feature of CAT is that it is possible—though not necessarily easy—to establish exposure parameters or decision rules that control the selection of test items (Thissen & Mislevy, 1990). By incorporating these item exposure controls, each test taker could be presented with a completely unique set of test items. Clearly, this capability greatly reduces, if not eliminates, threats to test security.

Additionally, as noted by Green (1983), CAT enhances security because the computer contains the item pool, rather than just the specific subset of items that will comprise the actual test. This makes it very difficult for test takers to spuriously improve their scores by learning a few items. Still, every effort should be made to ensure that the item pool is secure.

Summary and Recommendations

It is very likely that a candidate's test score will increase upon retesting, particularly if the same test is administered on each occasion. This gain, however, cannot be attributed exclusively to growth in a candidate's knowledge or skill base;

part of this gain may simply be due to a candidate's previous familiarity with the test—a practice effect. One potential consequence of this is granting a license to someone who does not possess the knowledge and skills necessary to protect the public's health, safety, and welfare. Licensing boards must, therefore, try to minimize the effects of practice on licensure test performance. The following suggestions are offered to help boards mitigate the effects of practice:

- Use alternate forms. Alternate or spiralled test forms help safeguard against item-specific practice effects. A candidate's recall of the item from a previous administration cannot come into play because the same items are not included on the alternate forms.
- Extend the time between test administrations. Few studies have examined the stability of practice effects over long periods of time. Nevertheless, a reasonable expectation is that the effects of practice will be less pronounced when the interval between test administrations increases.
- Use non-multiple-choice items. To our knowledge, no research has been conducted examining the effects of practice on non-multiple-choice items. The study by Powers et al. (1993), indicates, however, that coaching (viewed as an upper limit on practice) does not significantly affect constructed-response items. The use of non-multiple-choice items to reduce the effects of practice should be explored.
- Use computerized adaptive testing. The allure of computerized adaptive testing is its capacity to develop, on the spot, unique forms of a licensure test, thus potentially eliminating the effects of practice. The technical requirements to see this to fruition are not trivial, however. As work continues in this area, the use of this testing option should become more feasible.

COACHING

The preceding discussions of testing accommodations and practice effects treated broad questions of fairness in the context of high-stakes licensure tests. The question of fairness arises again on the issue of coaching, a technique some have embraced in attempts to improve their test scores.

The term "coaching" covers a wide variety of test-preparation activities that some view in a negative light. Clearly, research on the effects of coaching deserves the same thoughtful discussion we have given studies dealing with testing accommodations and practice effects—and for many of the same reasons, as we shall see.

Although coaching in athletics is generally thought a positive and often necessary activity, coaching for tests sometimes has negative connotations, in that test coaching is perceived as an illicit or, at least, nebulously inappropriate activity (Cole, 1982). Nevertheless, test coaching is a widespread enterprise. Many high schools provide in-class, instructional preparation for college entrance examinations. An ever-growing commercial industry provides test preparation courses for college, graduate school, and professional examinations. Test preparation books

and software packages are available in almost every library and bookstore in the country.

As Powers (1993a) notes, test preparation today is most often associated with high-stakes tests. These include assessments that are used either to select students for undergraduate and graduate study; to determine that they have demonstrated sufficient knowledge and/or skills to leave formal instructional settings; or to certify or license them in their professional careers. In some situations, such as those in which tests are used for accountability, both educators and administrators often have an interest, albeit somewhat vested, in making sure students are well prepared to take tests (Powers, 1993a).

Test publishers are also paying more attention to preparation. They are taking more responsibility to ensure that all test candidates are on as nearly equal ground as possible with respect to the methods required for good test taking. As Powers (1993a) notes, their rationale is straightforward.

> To be valid indicators, test scores should reflect the substance of the assessment much more than the method of assessment. Simply put, tests should reflect more than just the ability to take tests. (p. 2)

What is Coaching? Anastasi (1981) distinguishes three broad types of test preparation and discusses their implications for test taking. The first, test-taking orientation, entails test practice, which may help instill confidence and relieve anxiety by providing opportunities to learn appropriate test-taking strategies. The rationale for this intervention is that it can put all examinees on an equal footing with respect to their sophistication about test taking. A second type of preparation involves instruction in broad cognitive skills designed to develop intellectual skills and problem-solving strategies that may have broad application. This intervention, which might best be termed education, should improve both test scores and criterion performance. The third type of intervention concentrates on the specific knowledge and skills covered by the test, rather than more broadly on the larger domain that the test is intended to reflect. This type of intervention, according to Anastasi, is coaching. Bond (1989) espouses a similar definition of coaching. In his view, any instruction given primarily to increase test scores on a particular examination and only incidentally to improve the more general skills that the test is designed to measure can be considered coaching. Other writers (e.g., Slack & Porter, 1980) have argued that coaching includes any intervention, including full-time instruction for periods of 6 months or more, that results in improved test scores. The dictionary also presents a broadly inclusive definition, "to train intensively by instruction, demonstration, and practice" (Webster, 1974, p. 213). For the purposes of this paper, we will adopt Messick's (1982) definition: Coaching is "any intervention procedure specifically undertaken to improve test scores, whether by improving the skills measured by the test or by improving the skills for taking the test, or both" (p. 70). Therefore, "coaching'' and ''test preparation activities'' will be used interchangeably in this chapter.

A list of test preparation activities is provided by Cole (1982). She lists the following six components of test preparation: (1) supplying correct answers

(cheating), (2) taking the test for practice, (3) maximizing motivation, (4) optimizing test anxiety, (5) instructing in test wiseness, and (6) instructing in test content.

Components 5 and 6 are further delineated. Instruction in test wiseness includes: (a) general test-wiseness instruction (being careful, following directions, using good guessing strategies); (b) instruction in identifying test construction flaws and cues; and (c) use of special strategies for a novel or complex question format. Test wiseness may be generally defined as "a subject's capacity to utilize the characteristics and formats of the test and/or test-taking situation to receive a high score" (Millman, Bishop, & Ebel, 1965, p. 707). Instruction in test content, Component 6 in Cole's list, also has three subcomponents: (a) instruction in areas related to the interpretation of scores (the content domain for an achievement measure, the ability being measured, requisite skills or knowledge for eventual success for an admissions or selection measure); (b) review of previous instruction in areas related to score interpretation; and (c) instruction in test-specific content unrelated to score interpretation.

Test Preparation and Validity. Test preparation raises questions regarding test validity. Each individual enters the testing situation with his or her own assortment of skills, knowledge, experience, and characteristics. The testing situation is intended to produce a *sample* of performance in order to infer something more general about the individual. The extent to which such samples of performance (i.e., test scores) lead to correct interpretations of the more general domain is validity. Test preparation activities can have different effects on validity. These activities can give rise to three broad outcomes: (a) criterion performance overprediction, (b) predictor noise reduction, and (c) criterion and predictor performance gains. The particular outcome is entirely dependent on the nature of the test preparation activity.

Criterion Performance Overprediction. Efforts to improve the performance sample in the test without concomitant energy on the more general domain being measured poses a serious threat to validity. If coaching raises test performance above ability levels, then scores cannot be interpreted as accurate measures of ability. In Cole's (1982) scheme, the first component, supplying correct answers (cheating), would lead to this negative outcome. The result is that the test candidate may move from what Bond (1989) terms a "valid rejection" category to a "false acceptance" category. What is learned for the test is not transferred to the criterion; criterion performance is overpredicted as a result.

Cheating, once confined to glancing at your neighbor's bubble sheet, has advanced significantly in recent years. Technology and ingenuity have combined to present formidable challenges to test security. Testing companies and agencies regularly expose schemes involving paid and unpaid imposters. Some paid imposters may be hired (at additional cost) to resemble a candidate. The information age has also aided and abetted the cottage industry of test cheating. Facsimile machines, high-speed transoceanic and transcontinental flights, and tape recorders have been exposed recently as tools used to circumvent the testing process.

Subcomponent 5b, instruction in identifying test construction flaws and cues, may also result in test scores that overestimate knowledge and skills. Conse-

quently, test developers should be careful to screen assembled tests for item cue and overlap. Similarly, Subcomponent 6c represents instruction in content that is important to know in order to do well on the test, but is unrelated to criterion performance. For many kinds of test content, it is difficult to imagine an example of this subcomponent. Some item types, however, such as verbal analogies are rarely seen outside a test. Specific instruction in verbal analogies might improve test performance, but probably would not result in an increase in a student's academic performance. A licensure test, assuming a good job analysis and a specification plan that closely matches test content to job requirements, should be less susceptible to this type of overprediction.

Predictor Noise Reduction. Components 2–4 in Cole's scheme may also affect test validity. Unlike techniques that lead to overprediction of criterion performance, preparation activities that include test practice and that promote individual motivation and optimize test anxiety should allow candidates to better show their true ability. These activities would seem to be in the best interest of the test candidates, the test publishers, and all users of test scores. Further, Subcomponents 5a, instruction on general test wiseness, and 5c, use of special strategies for novel or complex question formats, might also enable the test-anxious student to be more relaxed and efficient during the test. In this instance, test performance would be improved and should be a more accurate reflection of ability. Such instruction does not enable students to achieve scores that overestimate their true level of knowledge and skills. Rather, it reduces the chances of underperforming (Jones, 1986). Such test preparation might result in a candidate moving from a "false rejection" category to a "valid acceptance" classification, an indisputably positive outcome (Bond, 1989).

If, however, test preparation of this type is only available to some candidates, the differences in the extent to which near-maximal performance is achieved could affect the validity of interpretation of the scores (Cole, 1982). This situation has social implications as well. If candidates who can afford special test preparation and coaching schools gain an advantage on admissions and professional licensure or certification tests, then testing could contribute to a sharper economic stratification in society. This result runs counter to testing's traditional goal of offering opportunity to the most capable regardless of economic background. For a test like the College Board SAT, for which there are a large number of books, software packages, and special preparation programs, the potential for unfairness is significant. As of 1988, there were at least 20 books and 30 software packages designed specifically to help students prepare for this single test (Powers, 1988). The greatest threat to equity, however, comes from the differential availability of formal commercially offered coaching programs. These programs may require substantial investments of time (up to and exceeding 40 hours of in-class instruction plus a large amount of time for homework and practice) and money. As these programs generally guarantee substantial score improvements but are not accessible to all, the public perception is that unfairness exists (Powers, 1993a). This persists despite the fact that the coaching-school claims for large score gains on the SAT have not been substantiated (cf. Messick & Jungeblut, 1981; DerSimonian & Laird, 1983; Kulick, Bangert-Drowns, & Kulick, 1984; Becker, 1990, Powers, 1993b).

Some authors (e.g., Downey, 1977; Sarnacki, 1979, 1990) have suggested general instruction in test wiseness for all test takers in order to attempt to eliminate or minimize the test-wiseness variable. Test publishers and agencies seem to have heeded this advice. Candidate information bulletins containing test descriptions, general test-taking strategies, and sample questions are generally provided to test candidates well in advance of the test date. More detailed information that might include the test specifications or body of knowledge, practice tests, and disclosed tests are often provided as well, particularly for tests with relatively large volumes. It should be noted, however, that Stricker (1982) found no discernible influence from disclosed tests on the SAT.

Criterion and Predictor Performance Gains. A third situation in which coaching can affect validity applies to strategies that focus on the criterion domain. Subcomponent 6a, instruction in areas related to the interpretation of scores, is such a strategy. For professional certification, 6a involves instruction in the knowledge and skills required for practicing the profession. For standardized achievement testing, it involves instruction in the knowledge and skills taught in the classroom. For admissions and selection, 6a involves instruction in the requisite knowledge and skills required for college, graduate, or professional education or a job (Cole, 1982). This strategy is a legitimate and defensible form of coaching, as it would raise both the level of test performance and facility within the domain being assessed. Assuming the test measures knowledge and skills that take time to acquire, this strategy must be associated with a reasonably long-term educational effort. In contrast, reviewing previously learned material relevant to the criterion, Subcomponent 6b requires much less time, but can also lead to performance improvements on both predictor and criterion. For the borderline candidate, coaching activities that focus on the criterion domain should have the effect of moving the student from the "valid rejection" category to the "valid acceptance" category.

The sole difficulty with strategies that focus on the criterion domain is that they rely heavily on the test as an authentic and representative sampling of that domain. If the test misses the mark, then well-prepared candidates will be underpredicted. They will be moved from a "valid acceptance" to a "false rejection" classification. This is one reason job analysis is critical for licensure and certification testing.

Coaching and New Forms of Assessment. Assessment is currently undergoing some very dramatic changes. The trends toward an emphasis on performance assessment, authentic assessment, computer-based assessment, and constructed-response item types will, no doubt, have ramifications for test coaching. It is too early to tell, however, just what the effects will be. Certainly, some measures might be less susceptible to illicit coaching, whereas others might be more so. For example, short-answer, open-ended items presented and scored by computer should resist coachability. Computerized adaptive tests, which by matching items with ability estimates are shorter and therefore expose fewer items, should also be less vulnerable to various forms of cheating (see Chapter 12).

The coachability of performance assessments is uncertain, but will likely depend upon fidelity of simulation and sufficiency of instruction. An oft-spoken

criticism about standardized testing—that teachers end up "teaching to the test"—ironically seems relevant here. The argument against teaching to the test seems based on the assumption that the test is not worthy of teaching to; that the educational experience will have little positive outcome as the test does not reflect the real world. In apparent contrast, performance assessments, which are supposed to simulate important criterion behavior, should be worthy of instruction. Therefore, if the assessment has high fidelity and the instruction is comprehensive, then the assessment should predict and the instruction should transfer to the criterion.

Recommendations for Licensure and Certification Programs. What is the relevance of coaching for "high-stakes" licensure and certification programs? What can be done to reduce threats to validity? A brief list of recommendations follows:

- Understand the criterion domain so that the test is a true reflection of the profession in question. Any test preparation activities that focus on the test content should thus provide at least some relevant education. The best way to maintain a strong link between the test and the profession is through periodic job analysis, followed by systematic test development.
- Provide adequate test information to all candidates in advance of the test. To help ensure candidates are on the same level playing ground, adequate test information should be provided in a candidate information bulletin. The bulletin should include: an overall description of the test, test-taking strategies, policy information about guessing and other relevant scoring issues, sample test items (particularly if they are at all novel), and information about the specifications for the test.
- Promote worthwhile educational activity. Licensure and certification programs might undertake several activities to promote education via testing. They could promote education by providing lists of reference texts and articles, publishing study manuals, and conducting review courses, for example.
- Maintain secure tests. Test security is the only safeguard against cheating. The initial stages of test development through test scoring and reporting must be secured. Further, item pools must be replenished on a regular basis.
- Review test items and forms for possible test-construction flaws. Test items should be carefully screened for flaws that might cue the correct answer. Assembled tests should be reviewed to minimize item overlap.
- Conduct item analysis. Even careful review may not identify all possible test-construction flaws prior to administration. Item analysis, however, may identify misbehaving items that may be flawed.

CONCLUSION

Testing special populations, testing repeaters, and coaching all have implications that can affect the validity and fairness of licensing examinations. This chapter has presented some important issues related to each of these topics as well as their psychometric implications. In addition, we have provided advice licensing

boards can consider when establishing or reviewing related policy issues. It is important that policies encourage equal access and fairness, and do so in a way that assures confidence in licensing as one way of protecting the public from incompetent practitioners.

Our review of the literature indicated that very little research on these topics was conducted within the context of licensure testing. This requires that boards set policy based on information and research findings from other contexts. Researchers and licensing boards must conduct studies to guide board policies on these topics.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

American Psychological Association, Division of Evaluation, Measurement, and Statistics. (1993, January). Psychometric and assessment issues raised by the Americans with Disabilities Act (ADA). *The Score, 15*(4), 1-15.

Anastasi, A. (1981). Coaching, test sophistication, and developed abilities. *American Psychologist, 36*(10), 1086-1093.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Becker, B. J. (1990). Coaching for the Scholastic Aptitude Test: Further synthesis and appraisal. *Review of Educational Research, 60*, 373-417.

Bennett, R. E., Ragosta, M., & Stricker, L. (1984). The test performance of handicapped people (Research Rep. No. 84-32). Princeton, NJ: Educational Testing Service.

Bond, L. (1989). The effects of special preparation on measures of scholastic ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 429-444). New York: Macmillan.

Burns, R. L. (1985). Guidelines for developing and using licensure tests. In J. C. Fortune (Ed.), *Understanding testing in occupational licensing* (pp. 15-44). San Francisco, CA: Jossey-Bass.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana: University of Illinois Press.

Cole, N. (1982). The implications of coaching for ability testing. In A. Wigdor & W. R. Gardner (Eds.), *Ability testing; Use consequences and controversies, Part 2: Documentation sections* (pp. 389-414). Washington, DC: National Academy Press.

DerSimonian, R., & Laird, N. M. (1983). Evaluating the effect of coaching on SAT scores: A meta-analysis. *Harvard Educational Review, 53*, 1-15.

Downey, G. W. (1977, January). Is it time we started teaching children how to take tests? *The American School Board Journal, 164*, 26-31.

Dunlap, K., & Snyder, A. (1920). Practice effects in intelligence tests. *Journal of Experimental Psychology, 3*, 396-403.

ETS Committee for Testing Handicapped People. (1988). *Sourcebook for testing handicapped examinees.* Princeton, NJ: Educational Testing Service.

Equal Employment Opportunity Commission and the U.S. Department of Justice. (Oct. 1991). *Americans with Disabilities Act Handbook.* Washington, DC: Equal Opportunity Commission and the U.S. Department of Justice.

Green, B. F. (1983). The promise of tailored tests. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement* (pp. 69-80). Hillsdale, NJ: Erlbaum.

Holland, P. W. (1981). *Section pre-equating the Graduate Record Examination* (Program Statistics Research Tech. Rep. No. 81-13). Princeton, NJ: Educational Testing Service.

Hopkins, K. D., Stanley, J. C., & Hopkins, B. R. (1990). *Educational and psychological measurement and evaluation* (7th ed.). Angled Cliffs, NJ: Prentice Hall.

Impara, J. C., & Stoker, H. W. (1985). Determining reliability and validity of licensure examinations. In J.C. Fortune (Ed.), *Understanding testing in occupational licensing* (pp. 65-86). San Francisco, CA: Jossey-Bass.

Jones, R. F. (1986). The effect of commercial coaching courses on performance on the MCAT. *Journal of Medical Education, 61*, 273-284.

Kane, M. A. (1982). The validity of licensure examinations. *American Psychologist, 37.* (8), 911-918.

Kulik, J. A., Bangert-Drowns, R. L., & Kulik, C. C. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin, 95*, 179-188.

Kulik, J. A., Kulik, C. C., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal, 21*, 435-447.

Laing, J., & Farmer, M. (1984). *Use of the ACT assessment by examinees with disabilities* (ACT Research Rep. No. 84). Iowa City, IA: American College Testing Publications.

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research, 55*, 387-413.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Madaus, G., & Mehrens, W. A. (1990). Conventional tests for licensure. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers* (pp. 257-277). Newbury Park, CA: Sage.

Messick, S. (1982). Issues of effectiveness and equity in the coaching controversy: Implications for educational and testing practice. *Educational Psychologist, 17*, 67-91.

Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin, 95*, 191-216.

Millman, J., Bishop, C. H., & Ebel, R. L. (1965). Analysis of test-wiseness. *Educational and Psychological Measurement, 25*, 707-726.

Phillips, S. E. (1993). Testing condition accommodations for disabled students. *West's Education Law Quarterly, 2,* (1), 366-389.

Powers, D. E. (1986). Relations of test item characteristics to test preparation/ test practice effects: A quantitative summary. *Psychological Bulletin, 100*, 67-77.

Powers, D. E. (1988). *Preparing for the SAT: A survey of programs and resources* (College Board Rep. No. 88-7 and ETS Research Report No. 88-40). New York: College Entrance Examination Board.

Powers, D. E. (1993a). Coaching for tests and examinations (Research Memorandum No. 93-7). Princeton, NJ: Educational Testing Service.

Powers, D. E. (1993b). Coaching for the SAT: A summary of the summaries and an update. *Educational Measurement: Issues and Practice, 12*, 24-39.

Powers, D. E., Fowles, M. E., & Farnum, M. (1993). Prepublishing the topics for a test of writing skills: A small-scale simulation. *Applied Measurement in Education, 6*, 119-135.

Ragosta, M., & Kaplan, B. A. (1986). *A survey of handicapped students taking special test administrations of the SAT and GRE* (Research Rep. No. 86-5). Princeton, NJ: Educational Testing Service.

Richardson, F., & Robinson, E. S. (1921). Effects of practice upon the scores and predictive validity of the alpha intelligence examination. *Journal of Experimental Psychology, 4*, 300-317.

Rock, D., & Werts, C. (1980). *An analysis of time-related score increments and/or decrements for GRE repeaters across ability and sex groups* (GRE Board Rep. GREB No. 77-9R). Princeton, NJ: Educational Testing Service.

Sarnacki, R. E. (1979). An examination of test-wiseness in the cognitive domain. *Review of Educational Research, 49*, 252-279.

Sarnacki, R. E. (1990). Test-wiseness. In H. J. Walbert & G. D. Haertel (Eds.), *The international encyclopedia of educational evaluation* (1st ed.; pp. 124-125). New York: Pergamon Press.

Schmitt, K. (1991). Testing across the nation. *Clear Exam Review, 2*(1), 4-6.

Sherman, S., & Robinson, N. (Eds.) (1982). *Ability testing of handicapped people: Dilemma for government, science, and the public.* Washington, DC: National Academy Press.

Shimberg, B. (1981). Testing for licensure and certification. *American Psychologist, 36.* (10) 138-1146.

Shimberg, B. (1985). Overview of professional and occupational licensing. In. J.C. Fortune (Ed.), *Understanding testing in occupational licensing* (pp. 1-14). San Francisco, CA: Jossey-Bass.

Slack, W. V., & Porter, D. (1980). The Scholastic Aptitude Test: A critical appraisal. *Harvard Educational Review, 50*, 154-175.

Stricker, L. J. (1982). *Test disclosure and retest performance on the Scholastic Aptitude Test* (College Board Rep. No. 82-7). New York: College Entrance Examination Board.

Thissen, D., & Mislevy, R. J. (1990). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 103-135). Hillsdale, NJ: Erlbaum.

Thorndike, E. L. (1922). Practice effects in intelligence tests. *Journal of Experimental Psychology, 5*, 101-107.

Wainer, H. (1990). Introduction and history. In H. Wainer (Ed.). *Computerized adaptive testing: A primer* (pp. 1-21). Hillsdale, NJ: Erlbaum.

Webster's new collegiate dictionary. (1974). Springfield, MA: G. & C. Merriam.

Weiss, R. A. (1961). *The effects of practicing a test: A review of the literature* (ETS Rep. No. RM-61-12). Princeton, NJ: Educational Testing Service.

Willingham, W. W., Ragosta, M., Bennett, R. E., Braun, H., Rock, P. A., & Powers, D. E. (1988). *Testing handicapped people.* Needham Heights, MA: Allyn and Bacon.

Wing, H. (1980). Practice effects with traditional mental test items. *Applied Psychological Measurement, 4,* 141-155.