6. The Future of Testing for Licensure and Certification Examinations

Michael T. Kane

The American College Testing Program

Follow this and additional works at: http://digitalcommons.unl.edu/burosfuturetesting

http://digitalcommons.unl.edu/burosfuturetesting/7

This Article is brought to you for free and open access by the Buros-Nebraska Series on Measurement and Testing at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in The Future of Testing by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.
Licensure and certification examinations constitute a major use of tests in the United States, and since licensure and certification provide obvious benefits to the persons licensed or certified and to the public, their use is not likely to decrease. Rather, the usage of such examinations to document competence is likely to continue to increase, although perhaps more slowly than it has in the recent past.

Critics of licensure have argued that licensure tends to benefit the licensed profession more than the public and that the benefits of licensure to the public do not always justify the costs (e.g., see Hogan, 1979; Williamson, 1976). However, the weight of criticism of licensure and certification tends to be that they do not provide sufficient protection rather than that protection is not needed. Furthermore, if licensure were eliminated in areas like the health professions, we would undoubtedly experience some increase in quackery, leading to demands for increased protection. Therefore, I expect that the criticisms of licensure and certification are more likely to change the social and legal context in which the various forms of credentialing operate than they are to decrease the extent of credentialing. In particular, the trend has been in the direction of greater public scrutiny of the activities of licensing and certifying bodies accompanied by demands for more public disclosure.

Both licensure and certification are credentials intended to document the possession of specialized knowledge and skills. Both forms of credentials confer on their holders certain privileges and responsibilities. They differ in the source of the credential and in the legal status of the credential.

Licensure is a state function and is usually administered by a state board with legal authority to regulate the practice of the profession. Although professional organizations have traditionally been involved in setting standards and nominat-
ing members of state licensure boards, ultimate authority rests with the state legislature. The laws vary from state to state and from profession to profession, but, in general, a license confers on its holder the right to use a title and to provide certain services that the licensure law makes illegal for nonlicensed persons to provide. It also subjects the licensed professional to regulation by the state licensing authority, often referred to as the “board.” The license is interpreted as indicating that its holder has the basic knowledge and skills required for safe and effective practice.

Voluntary certification programs are administered by professional organizations and do not generally have a formal legal status. The professional organization uses certification to recognize training and experience beyond the basic requirements for licensure. To the extent that certification works well, it provides the public with a basis for identifying individuals who are especially well qualified to handle certain kinds of specialized problems. Like licensure, certification can provide substantial benefits to both the practitioner and the public.

It is worth noting that there is considerable variation, and potential confusion, in the terminology used to describe various credentials. Although teachers are certified in most states, the requirements for teacher certification are state-imposed and mandatory for practice. In most school contexts, teacher certification is an example of “licensure” rather than “certification” as these terms are used here.

Although certification does not generally have a formal legal status, it is pervasive enough in medicine as to have significant legal and professional implications. Hospitals may not permit a practitioner who is not certified to provide services usually provided by board-certified practitioners. Furthermore, physicians who engaged in specialized activities, like major surgery, without being certified would expose themselves to punitive malpractice judgments. Therefore, in terms of the restrictions that a lack of certification imposes, some kinds of certification are effectively very similar to licensure.

Since the requirements for voluntary certification in terms of education, experience, and examinations are also quite similar to those for licensure, it will not be necessary for most of the discussion that follows to draw a sharp distinction between these two forms of credentials. Where the differences between the two types of credentials have a significant impact on the issue under discussion, I will try to make this clear, but for simplicity I will emphasize licensure, the more pervasive of the two kinds of credentials.

CURRENT STATUS

Although the specific requirements for licensure vary considerably across the professions and trades that are licensed and across the jurisdictions awarding these licensures, the general pattern is fairly consistent. The requirements typ-
6. FUTURE OF LICENSURE AND CERTIFICATION EXAMS

ica ll y involve four components: educational requirements, an examination, evidence of good character, and an ongoing policing function. Requirements for relicensure, including continuing education and/or retesting, have become more common, but they are still not the rule (Lowenthal, 1981, provides a recent overview of issues in continuing education for professionals). The main concern of this paper is, of course, the examinations, but some remarks about the other components are in order because they establish the context in which licensure decisions are made and therefore help to place in perspective the issues associated with examinations.

Educational Requirements

The educational requirements for licensure generally involve successful completion of an approved educational program. The requirements can be quite extensive and usually are quite detailed, often specifying, for example, the length of the program, particular courses to be included, etc.

These educational requirements have a significant impact on the interpretation of licensure examination results in at least two ways. First, they have implications for the specification of the content domain to be covered by the examination. Content which is viewed as providing a useful background for practice but having only an indirect or secondary impact on performance could reasonably be omitted from the licensure examination on the basis that this content is thoroughly taught and tested in the educational program; for example, the research methodology of a discipline might be given relatively little emphasis on the examination, assuming that it is covered in the educational program; for example, the research methodology of a discipline might be given relatively little emphasis on the examination, assuming that it is covered in the educational program. Furthermore, skills that are difficult to assess in a large-scale examination (e.g., performance skills like giving an injection or conducting an interview) are often omitted from the examination, based, at least in part, on the assumption that these skills are adequately documented by the educational program. These remarks suggest that the content of the licensure examination need not be the same as the content of the curriculum; indeed, it would be wasteful and counterproductive if one followed the other too closely. Nevertheless, we should expect a high degree of overlap between the content of professional school curricula and the content of licensure and certification examinations, since both presumably emphasize knowledge and skills that are viewed as needed for effective practice.

This leads to the second implication of the educational requirements. The existence of rigorous educational requirements provides some assurance that the persons taking the licensure examination are generally well prepared. Therefore, if the system as a whole is working well, the failure rate on a licensure examination that has rigorous education prerequisites should typically be relatively low; if the failure rate were very high, it would be reasonable to suspect that either the examination procedures or the educational programs are not functioning properly.
Of course, this point raises the question of whether the examinations, particularly licensure examinations, are needed at all. If the examination is assessing knowledge and skills that have already been assessed in the educational program, what function does the examination serve? In many cases (e.g., college teachers), professionals are allowed to practice on the basis of educational credentials without having to take any specific examinations.

For certification examinations, the relationship between the expected failure rate on the examination and the extensiveness and degree of rigor of the educational requirements that must be met before the examination can be taken is less clear, because certification is intended to document levels of competence that are often much higher than that required for licensure. Given the high level of skills expected for certification, even relatively lengthy educational preparation may not be viewed as providing strong assurance that most candidates are in fact qualified.

Both licensure and certification examinations can be effective in doing two things. First, the examination provides an additional check on the preparation of individual candidates for licensure. Given the inevitable variability of educational programs, some candidates with deficiencies in some areas of preparation are likely to graduate; the examinations provide evidence of practitioner competence, based on assessment procedures that are the same for all candidates. In a sense, the interpretation given to the examinations is Bayesian in that the educational record constitutes prior information indicating that most candidates for licensure are qualified. This view is reflected in the fact that graduates of foreign professional schools, for which less documentation of program content and quality is available, are often required to take a somewhat more extensive battery of examinations than is required of graduates of approved schools in the United States.

Second, the examination provides a measure of the variability in educational programs and helps to encourage consistency of standards across the programs within the state. It provides an external check on the quality of educational programs. The examination also provides an incentive for programs with disproportionately high failure rates to take steps designed to improve their graduates’ performance.

Evidence of Good Character and the Policing Function

The third and fourth types of requirements, evidence of “good character” and the ongoing policing function, are both designed to maintain ethical standards, although the policing function also covers questions of continuing competence. Evaluations of ethics, or “character,” raise obvious problems (including possible invasion of privacy), and the effectiveness of the policing functions of licensing boards has been criticized widely. Nevertheless, the existence of these special mechanisms for maintaining ethical standards is significant in that they reflect the general perception, which I think is well justified, that a written
examination does not provide an effective mechanism for evaluating ethics and related characteristics, like conscientiousness.

The importance of ethical considerations in determining the quality of professional practice is illustrated by a study of laboratory practice done at the Center for Disease Control in Atlanta (referenced by Williamson, 1976). For half of a set of blood samples, the laboratories knew their performance was being evaluated, and for the other half, the samples were simply sent in by a local physician with a patient's name on them. According to Williamson (1976), "A 4% deficiency rate occurred when the lab was aware it was being tested, whereas a 50% deficiency rate was found when the lab was not aware it was being assessed" (p. 24). The point is that the typical performance of both organizations and individuals falls short of what they would be capable of at their best, and the difference between typical performance and optimal performance is determined by the degree of care and effort that is devoted to an activity. Thus, conscientiousness and the larger issues of professional ethics are likely to be a major determinant of the quality of practice. Unfortunately, methods for assessing an individual's current level of ethics are rather weak, and our ability to predict future behavior is even weaker.

Probably the best available indication of a candidate's ethics is provided by the record of the candidate's performance in the required educational program. Because candidates have strong incentives to present themselves in the best possible light on a licensure or certification examination, candidates are likely to perform at levels close to their optimal levels of performance during the examination. Since the faculty in the education program have the opportunity to observe the candidates' performance in a variety of situations over a long period of time, they have a good opportunity to detect dishonesty, laziness, carelessness, etc., and this current indicator of ethics is probably the best predictor of future behavior. However, a policing function of some kind is needed to limit the negative consequences caused by practitioners who subsequently get into trouble (due, for example, to physical or mental illness, personal problems, financial difficulties, etc.), because we cannot predict such future developments with any accuracy.

It is worth noting that some licensure examinations also include items, or a separate test, on the ethical code for the profession. However, since such items cover knowledge and understanding of the rules of ethics rather than inclination to observe the rules, the requirement should be considered as part of the examination process rather than as part of the "good character" component.

Examinations

Having sketched some aspects of the other components that commonly occur in licensure procedures, we can turn to the central concern of this paper, the examinations themselves. In terms of format, most licensure examinations are written multiple-choice tests, although some also involve other forms of written
test (e.g., see Hubbard, 1971), and some have a performance component (e.g., see Kastinos & Livingston, 1979; Reed, 1978).

The content specifications for the examination, which provide an operational definition of the domain of knowledge and skills covered by the examination, are typically developed by members of the profession being licensed, with assistance from testing specialists. The items are also written and reviewed by members of the profession, who may or may not be the same persons responsible for the content specifications. In most cases, the items are reviewed and edited by testing specialists, and as part of this technical review, item analysis procedures are routinely employed.

**Determination of Passing Score.** After the examination has been prepared, a passing score is determined. There is wide variation in how this is done, but two general approaches can be identified: those based on the distribution of scores for some “norms” group, and those based on professional judgment. The norms-based methods, which are the more traditional, typically set the passing score at something like one or two standard deviations below the mean score of the norms group. An obvious disadvantage of the norms-based approach is that the performance of each candidate is judged relative to the performance of other candidates, those in the norms group, rather than being judged against the requirements of practice. Given the purpose of licensure and certification examinations, such relative standards do not seem to be appropriate.

The judgment-based standard setting procedures in common use are variants of those proposed by Nedelsky (1954) and Angoff (1971). In these procedures, experts review each item and determine a minimal pass level, or MPL, defined in terms of the probability that a minimally competent candidate would answer the item correctly. Presumably these estimates reflect the experts’ judgments about the importance of the content being tested and the difficulty of the item. (A method proposed by Ebel, 1972, explicitly incorporates judgments about importance and difficulty but is not as widely used.) The MPLs are then summed over items to obtain the passing score for the test. These methods have the advantage of being based on expert judgment and therefore of having a rational relationship to practice, but they have a number of problems of stability. Although they are intended to serve the same purpose, the methods tend to give different results (e.g., see Andrew & Hecht, 1976; Brennan & Lockwood, 1980; Shepard, 1980), and the consistency among raters using a given method is not especially high. Furthermore, neither the norms-based nor the judgment-based standard setting methods generate passing scores with an obvious interpretation in terms of practice requirements. This last issue is discussed in some detail later in this paper.

**Public Disclosure.** A trend of the recent past, which is likely to continue in the future, is greater public disclosure of the characteristics of the examinations as well as the details of licensure procedures in general. This trend has involved
such developments as the appointment of public members to licensure boards and sunset legislation mandating periodic legislative review of the work of licensure boards, as well as public disclosure of test items and test forms. Licensure tests have also been examined intensively by various state agencies (e.g., Werner, 1981). Certification examinations have been less subject to outside scrutiny but have also tended to move toward greater public disclosure. This trend should lead to more research on licensure and certification examinations and more thorough documentation of their characteristics, and therefore should facilitate informed debate.

As a final note on the current status of licensing and certification examinations, it is fair to say that the procedures used to develop the tests generally involve the traditional approach to developing standardized achievement tests; in some cases, they are classic examples of this methodology.

CRITICAL ISSUES

The central issue for licensure and certification examinations is validity, that is, the evidence for the interpretation of the results of the examination. Simply put, the question is: What can we justifiably infer about candidates for licensure or certification on the basis of their scores on the examination?

Validity and Utility

Validity is a fundamental concern and, as such, is related to a number of other issues, including the more general concern for the utility of specific forms of regulation embodied in certification and licensure. Presumably, the aim of such credentials is to protect the public, and the effectiveness of examinations in abetting this goal is based on two basic assumptions. First, it is assumed that the public needs protection, and that this need is sufficiently great that society should bear the considerable expense imposed by licensure and certification procedures. In medicine, where practitioners act relatively independently, where the public is generally not in a good position to judge the competence of practitioners, and where the consequences of incompetence can be severe, the protection provided by licensure and certification seems to be most justified. However, it should be noted that even in the case of medicine, the argument has been made that licensure serves the interests of the profession more than it serves the interests of the public (e.g., see Gross, 1978). In other fields (e.g., cosmetology), the need for relatively expensive forms of protection, like licensure, is less clear, but in any case, the public through the political process must decide how much protection/regulation it wants to buy and how much of this protection is achieved most effectively by the regulation of individuals. In some cases, it is clearly more efficient for the state to regulate organizations as is done via safety regulations in industry.
Second, given that protection from incompetent practitioners is needed, it is assumed that licensure or certification, and in particular the examinations required for these credentials, afford the desired protection. This assumption, and therefore the utility of an examination as part of the overall process, depends on the interpretation given to the results of the examination and on the evidence for the proposed interpretation. In particular, when evaluating a licensure or certification examination, the case must be made that those who pass the examination are more likely to be safe and effective practitioners than those who fail the examination. This case will rest on the evidence for the validity of the examination and, as a related issue, on the justification for the procedures used to establish the passing score.

Trade-off Between Utility and Validity. In a sense, there is a trade-off between the utility, or import, of the type of interpretation assigned to examination results and the ease with which the interpretation can be validated (see Kane, 1982b). If the interpretation given to the examination involves strong inferences, validation will be relatively difficult, but if validation were achieved, the examination would make a large contribution to the utility of the resulting decisions. More limited interpretations generally have less utility but are also easier to validate. For example, if a test consisting of questions about the ethical code for a profession were interpreted as a measure of knowledge of the ethical code, it would be relatively easy to validate and would have substantial utility for licensure decisions; under this interpretation, knowledge of the ethical code as reflected in performance on the examination is viewed as a necessary but not sufficient condition for observance of the code. If the test were given a stronger interpretation as a predictor of how ethical the candidate would be in practice, it could have great utility if validated but would be difficult, if not impossible, to validate adequately; under this interpretation, knowledge of the ethical code is viewed as a sufficient condition for observance of the code.

Bias. Closely related to the issue of validity is the issue of bias. To the extent that candidates who have acquired the skills needed for practice fail an examination because of irrelevant factors such as race, sex, or the existence of a handicap that would not interfere with effective practice, the examination would not be valid. However, to the extent that the examination scores reflect candidates’ degree of preparedness for safe practice, they would not be considered biased even if they had adverse impact in the sense that the failure rate is higher in some groups than in others. The distinction between adverse impact and bias is embedded in the Uniform Guidelines for Employee Selection Procedures (Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice, 1978) used by federal agencies in enforcing civil rights legislation, and reflects the recognition that differential educational
experiences can lead to differential achievement. Since different failure rates for various groups, i.e., adverse impact, may result either from differential levels of preparation for the group or from bias in the examination, the Uniform Guidelines require evidence for validity in cases where significant adverse impact occurs.

Because they were developed to aid in the enforcement of federal civil rights legislation, the Uniform Guidelines do not require evidence for validity unless there is adverse impact against groups specifically protected by federal legislation. Furthermore, because of the special role of state government in our federal system, the Uniform Guidelines may not apply to licensure examinations. However, since the justification for the use of licensure and certification examinations depends on their interpretation, evidence for the validity of the proposed interpretation is needed to justify the use of such examinations, even if adverse impact is not found. Where adverse impact is found, the need for careful evaluation of the validity of the examinations is especially important.

Validating Licensure Examinations

Given that validity consists of the evidence supporting the proposed interpretation of examination scores, and that a candidate’s score on a licensure examination is interpreted as indicating the candidate’s readiness to practice safely and effectively, the required evidence for validity should establish a relationship between scores on the examination and readiness for practice. The issue, then, is the nature of this relationship and the evidence needed to establish that the intended relationship exists.

Since licensure laws are written by state legislatures and administered by state boards, the presumed relationship between scores on the examination and readiness for practice is determined by the legislature and by the state boards and therefore varies from state to state and from profession to profession. Similarly, for certification examinations, the presumed relationship between examination scores and performance in practice depends on the interpretation proposed by the certifying agency. The remarks that follow apply to the general goal of promoting “safe and effective” practice and would apply in general terms to most licensure and certification programs; these remarks represent a more fully developed discussion of suggestions made in Kane (1982a).

Validity consists of an argument for an interpretation of examination scores, and the evidence included in such an argument may take many forms. In most discussions of validity, the types of evidence are discussed under three headings: content validity, criterion validity, and construct validity. Content validity evidence supports the interpretation of test scores in terms of some domain of content and indicates that test scores reflect the degree of mastery of the content domain.
Evidence for criterion validity supports the interpretation of test scores as predictors of some criterion of interest. In the case of licensure and certification examinations, the criterion may be a measure of future performance (e.g., ratings of performance in practice), or it may be a score on some assessment of performance given at about the same time as the examination (e.g., a performance examination simulating some aspects of practice situations might be used to examine the validity of a multiple-choice examination). These two subclasses of criterion validity are called predictive validity and concurrent validity, respectively.

Construct validity supports the interpretation of test scores in terms of certain assumptions about what is being measured and indicates that the test scores reflect an attribute defined by the assumptions. The methods of construct validity that depart dramatically from the more traditional methods of content validity and criterion validity are most clearly applicable where the attribute being measured is implicitly defined by a theory. In such cases the assumptions used to generate validity evidence would be drawn from the theory (i.e., see Cronbach & Meehl, 1955). However, construct validity can also be viewed as subsuming content validity and criterion validity. In criterion validity, the assumption being investigated is that readiness for practice as measured by the examination is related, usually linearly, to subsequent performance in practice. In content validity, the assumption being tested is that the test measures knowledge of a domain that is important for performance in practice.

Because validity is associated with the interpretation of measurements (Cronbach, 1971), evidence that supports the intended interpretation of test scores supports claims for validity, and evidence that disagrees with the intended interpretation tends to refute claims for validity. As noted earlier, there are two common interpretations of the scores on licensure and certification examinations. First, they can be interpreted as providing predictions of an examinee’s future professional performance. Second, they can be interpreted as providing evidence of an examinee’s present competence on specific abilities that are needed for practice. The interpretation of licensure examinations as predictors of future professional performance suggests the use of predictive validity in evaluating licensure examinations. The interpretation in terms of abilities that are needed in practice suggests the use of content validity.

I have argued (Kane, 1982a) that content validity, considered broadly, would provide a more effective approach to investigating the validity of licensure examinations than can be provided by criterion validity. Manning (1978) has made a similar argument in discussing the legal aspects of validation for employment testing. However, before summarizing the reasons for this position, it is worth emphasizing that, to some extent, all three types of validity evidence are likely to occur in validating any test interpretation; the issue is one of emphasis rather than a choice between clearly separate and distinct approaches.
Criterion Validity—The Interpretation of Licensure Examination Scores as Predictors of Future Performance

The interpretation of licensure examination scores as predictors of future performance in practice is appealing because it implies a high degree of utility for the licensure process. To the extent that this interpretation does provide the justification for a licensure examination, arguments for validity would be based on empirical evidence indicating how well examination scores predict future performance, that is, on predictive validity (Hogan, 1979; Menges, 1975; Pottinger, 1979). Hecht (1979) has stated this position clearly:

It would appear to me that predictive criterion-related validation studies would be the type most closely fitting the expressed purpose of licensure exams, that of assuring minimal competency on the job for the protection of the public. Interest is with the criterion not yet obtainable at the time of testing. (p. 21)

Similarly, Andrew (1976) has emphasized criterion validity as the ultimate aim in validating certification examinations:

The challenge that faces us now should encourage us to get on with the business of establishing content validation for our examinations, and to turn our attention even more vigorously to the establishment of criterion-related validity for our certifying examinations. In doing so we must focus our attention on the development of techniques to assess criterion measurements of performance. (p. 46)

As illustrated by these quotations, predictive validity is often presented as the best approach for validating licensure and certification examinations, but this preference for predictive validity is not reflected in practice.

*The Criterion Problem.* The usefulness of predictive validity for licensure and certification examinations is limited greatly by the fact that criteria of proven validity are not available for licensure examinations. The development and validation of a criterion measure of professional performance presents fundamental conceptual problems as well as great practical difficulties, in part because practice requires a high level of professional judgment for effective performance. The distinction between good practice and poor practice is not clear-cut in most cases (e.g., see Strupp, Hadley, & Gomes-Schwartz, 1977), and the development of general measures of the quality of practice that are reliable, valid, and complete is probably not possible for most professions. Assumptions about the validity of the criterion are likely to be questionable at best, and to the extent that the validity of the criterion is questionable, any conclusions drawn from a predictive validity study would be questionable.
The seriousness of the criterion problem is illustrated by the experience of the National Board of Medical Examiners, as reported by Hubbard (1971). After 3 years of attempting to develop a reliable bedside evaluation using real clients, they found that when one observer rated a candidate in one situation and another observer rated the same candidate in a different situation, the interrater agreement was at the chance level. In another study, Hoffman (1977) found that an oral examination based on a physician's interaction with a client had low reliability because of variability in the assessments of performance from one situation to another. Where such results occur, one must conclude that the ratings are, to a large extent, measuring characteristics of the raters, the situations, or other contextual factors rather than the competence of the candidate.

**Technical Problems.** In addition to the criterion problem, there are two technical issues that limit the application of criterion validity to licensure and certification examinations. First, licensure is not intended to indicate readiness for a specific task or job, but rather for a wide range of activities in a variety of settings. A criterion validity study showing that a test predicts performance in one setting does not necessarily demonstrate that the test also predicts performance in other settings, and it is not clear whether evidence for criterion validity can be generalized from one setting to another (Cronbach, 1980a; Hunter, 1980). For a licensure examination, therefore, the logic of criterion validity could require not one validity study but a large number of validity studies—one for each of the settings in which those who are licensed might practice. For certification examinations, the range of practice situations is more restricted but is still quite broad.

A second technical problem is that the data needed to evaluate the predictive validity of a licensure examination are not generally available, because those who do not pass the examination are not allowed to practice. A licensure examination is not designed to predict varying degrees of expertise, but simply to distinguish those candidates who are prepared for practice from those who are not. The crucial question for a study of the predictive validity of a licensure examination is whether those who pass the examination are more likely to be safe and effective in practice that those who fail, and this question is not answered by a correlation coefficient based only on passing candidates. A more appropriate index of the predictive validity of a licensing examination would be a measure of the agreement between the pass/fail dichotomy on the licensure examination and a competent/incompetent dichotomy in subsequent practice; however, an index, like coefficient kappa (Cohen, 1960), that would address this issue cannot be estimated without having criterion scores for those who fail the licensure examination as well as for those who pass. Attempts to collect such data might be considered unethical (and probably illegal) in many professions.

This second technical problem does not apply with equal force to certification examinations, because, as this term is used here, certification is not a mandatory
requirement for practice in a specialized area. For example, physicians can treat children without being board-certified in pediatrics. Williamson (1976) discusses a number of studies that are relevant to the predictive validity of certification examinations. However, as noted earlier, the practice of a physician who is not certified in an area where certification is common is likely to be somewhat restricted by hospital policies, difficulties in getting malpractice insurance, etc. Also, individuals who choose to specialize in an area of practice will typically meet at least some of the requirements for certification in terms of education and experience even if they are not certified. Therefore, the differences in the scope of practice between certified and noncertified specialists and the overlap in credentials will make decisive studies of predictive validity difficult to implement even in the case of certification examinations.

A related issue that is not as serious as the two technical difficulties just described involves the determination of how strong the relationship between examination scores and the criterion measure must be in order to establish a reasonable case for criterion validity. In some cases, even a weak relationship (e.g., a relatively low correlation) might be sufficient to justify the use of a licensure or certification examination, since even a small increase in the average level of performance in a profession could yield major benefits for society. Furthermore, there are good reasons to expect that the relationship between scores on a licensure or certification examination and subsequent performance in practice would not be particularly strong. As indicated by Gonnella, Goran, Williamson and Cotsonas (1970), successful performance on an examination does not provide a guarantee that the examinee’s current level of performance in practice would be satisfactory.

Inferences to future performance are even more problematic since there are a number of factors (e.g., serious illness) that could have a major impact on the quality of future performance but cannot be predicted in advance. The interest in mandatory continuing education is based on the realization that practitioners vary in how well they maintain or enhance their skills after they enter practice. The requirement that small correlations be estimated with precision, combined with the intrinsic difficulties in conducting criterion validity studies for licensure and certification examinations, makes it unlikely that such a study would yield dependable results.

The Interpretation of Licensure Examination Scores as Measures of Critical Abilities

The severe problems associated with predictive validity can be avoided by interpreting the test scores in terms of a domain of knowledge and skills required for practice. The knowledge and skills included in the domain are assumed to be “critical” in that they are necessary, although not sufficient, for effective performance in practice. Abilities are considered critical to the extent that their absence
would be a serious limitation in the practice of the profession. The critical abilities for a profession typically include cognitive abilities involving knowledge and the ability to apply knowledge, as well as psychomotor skills involving the ability to apply various skills for clients.

In interpreting licensure and certification examinations in terms of critical abilities, the connection between test performance and performance in practice involves two steps. First, the test scores are interpreted as indicating overall level of proficiency in a domain of critical abilities, and second, some level of proficiency in the domain is viewed as necessary for effective performance in practice.

**Abilities as Necessary but Not Sufficient Requirements.** The fact that skills that are necessary for effective performance do not generally guarantee effective performance is illustrated by the study on the treatment of urinary tract infections mentioned earlier (Gonnella et al., 1970). In this study, the performance of patient care teams in detecting and treating urinary tract infections was evaluated by a review of clinic charts, and the team members were given a 50-item multiple-choice examination and a simulated clinical problem dealing with urinary tract infection. The authors concluded that:

> In the comparison of knowledge and performance major discrepancies were found in our study. It is disturbing to learn that on an examination the students and physicians indicate that a history of catheterization, nephrolithiasis, past treatment of urinary tract infection, hypertension, and diabetes mellitus are critical data but in an actual treatment situation either fail to ask these questions or fail to follow through once the information has been obtained. (p. 2043)

The possession of critical knowledge and skills does not guarantee that the knowledge or skills will be used effectively. The clinic situation, in which the physician deals with multiple patients, interacts with many other professional staff, and must wait for lab results for hours or days, is quite different from the examination situation, in which the facts are presented in an orderly fashion and there are no distractions. However, it is safe to assume that persons who do not possess the required knowledge and skills will not be likely to make use of them. Thus, the critical abilities are necessary but not sufficient requirements for effective practice.

**Critical Abilities and the Department of Learning.** What kind of abilities should be considered critical abilities for a profession? The *American College Dictionary* defines a profession as a "vocation requiring knowledge of some department of learning or science." Presumably many of the critical abilities will be included in the department of learning or science associated with the profession. The abilities may be quite general (e.g., communication skills) or quite specific (e.g., the ability to carry out a particular procedure). Including a particu-
lar ability in a licensure or certification examination would be justified by evidence connecting the ability to client outcomes, and typically this evidence would be drawn from the department of learning for the profession. The inclusion of some abilities is based on empirical evidence (e.g., ability to carry out medical procedures that are based on clinical trials). In other cases, abilities are justified by logical analysis and by procedural rules (e.g., in law). Specification of test content in terms of critical abilities does not require an exhaustive listing of the abilities required for practice, but each ability should be clearly related to practice. Where certification follows basic licensure, the critical abilities for certification include all those required for licensure and, in addition, include specialized knowledge and skills in the area of certification.

**Structure of Validity Arguments.** The structure of validity arguments involving the critical ability approach is quite simple, including two premises and a conclusion. The first premise states that, because the critical abilities are necessary for effective performance, individuals who lack the critical abilities to a substantial degree will not be able to perform adequately in practice. The second premise states that individuals who have low scores on the examination lack the critical abilities to a substantial degree. The conclusion which follows from these two premises states that individuals who have low scores on the examination will not be able to perform adequately in practice.

Although the structure of the argument is simple, the development of such arguments in specific cases is not simple because it requires substantial evidence for the two premises. The second premise involves issues usually considered under the label of "content validity" (i.e., relationship between test and domain) and issues of standard setting (i.e., what does it mean to say that individuals lack the critical abilities "to a substantial degree"?).

The first premise assumes a relationship between the critical abilities and performance in practice. If it were necessary to start from scratch, justification for the relationship between critical abilities and performance in practice could be an enormous undertaking; a large-scale study might be required to establish a relationship between a particular intervention (e.g., polio vaccination) and the quality of professional practice defined in terms of client outcomes (e.g., incidence of polio). Fortunately, it is not necessary to start from scratch. The department of learning for a profession often includes a large body of data on the relationship between abilities and outcomes. In fact, much of the research effort included in the relevant department of learning can be interpreted as an attempt to identify critical elements in the practice of the profession. To the extent that this research has been replicated and subjected to careful review without being refuted, we have a reasonable basis for confidence in the results.

It is undoubtedly the case that the department of learning for every profession is incomplete, and in some respects incorrect, but for most professions it does represent a substantial body of knowledge about the critical requirements for
practice. Therefore, the department of learning establishes a connection between various critical abilities and the quality of practice, and provides the justification for demanding some level of mastery of the critical abilities as a requirement for licensure or certification. As a result, a validation strategy based on critical abilities can concentrate on the second premise, showing that the examination results can be interpreted as indices of the level of proficiency in the required critical abilities.

Of course, the critical abilities approach to validation has its problems and, like criterion validation, is no panacea. The departments of learning are often large and are seldom organized in a way that is appropriate for test development. Therefore, expert judgment is involved in organizing the department of learning for test development purposes (i.e., defining a table of specifications for the test). This effort requires evaluation of the relative importance of various parts of the domain, and such judgments are always fallible. Empirical studies of patterns of practice can help to evaluate the relative importance of different abilities, and therefore provide a useful check on these judgments.

Combining Validation Strategies

The critical abilities approach incorporates aspects of content validity, criterion validity, and construct validity. The evidence supporting the interpretation of test scores in terms of a domain of critical abilities would incorporate many elements of content validity. Several of the issues that arise in this context are discussed in the next section, labeled Changes Needed, and in comments on empirical job analyses that appear later in the paper.

The evidence relating critical abilities to client outcomes can be interpreted as providing indirect criterion validation of the licensure examination. A predictive validity study seeks to determine the relationship between performance on a test and some criterion of future performance for each individual, while the critical abilities approach depends on the relationship between an ability and client outcomes averaged over large numbers of professionals and clients (i.e., in clinical trials) or on rational analysis (as in law and some aspects of teaching). Such studies are likely to provide the most accurate analysis available of the importance of various abilities for professional practice.

In a sense, the difference between the predictive validity approach and the critical abilities approach is that the predictive validity approach is almost purely empirical, while the critical abilities approach depends heavily on both the theoretical and empirical content of the department of learning associated with the profession. Studies of predictive validity draw on the "department of learning" in defining the criterion but usually take the examination as a given and proceed to evaluate the empirical relationship between examination scores and criterion scores.
The critical abilities approach makes more extensive use of the theory and the accumulated body of empirical findings in the department of learning, which it uses to define an appropriate content domain. The domain definition is subject to challenge, and empirical job analyses can be employed to investigate some possible challenges. The examination designed to measure mastery of the domain is also subject to challenges of various kinds, and the discussion in the next section will elaborate on the nature of some of the possible challenges and the steps that can be taken to evaluate such challenges.

In its emphasis on the department of learning and the empirical testing of assumptions based on this body of knowledge, the critical abilities approach requires arguments/analyses that are more complicated than those typically employed in studies of criterion validity and content validity. This more general form of validity evidence can be viewed as an example of construct validity, where the construct at issue, professional competence, is defined in terms of the network of theoretical and empirical relationships incorporated in the department of learning.

Testing Standards and Guidelines
In part because of their increasing visibility, licensure and certification examinations have been discussed explicitly in several recent documents containing standards or guidelines for test preparation and use. The most prominent of such documents is the Joint Technical Standards for Educational and Psychological Testing, published in draft form in February of 1984 by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education.

Joint Technical Standards. Chapter 13 of the draft standards (AERA, APA, & NCME, 1984) is devoted to standards for licensure and certification examinations. The introduction to chapter 13 acknowledges the difficulties in conducting sound predictive validity studies for licensure and certification and suggests that:

The difficulty in conducting criterion-related validation studies does not, however, lessen the importance of validity, which remains a central concern. The test user should develop the evidential basis to support the particular use. For licensure and certification, however, primary reliance must usually be placed on content evidence supplemented by evidence of the appropriateness of the construct being measured. (p. 13–2)

This suggestion, combined with the first standard in chapter 13, quoted below, reflects the basic rationale for a validation strategy based on critical abilities.

Standard 13.1. The content domain to be covered by the test should be clearly defined and explained in terms of the importance of the content for competent
performance in the occupation. A rationale should be provided to support a claim that the knowledge or skills being assessed are required for competent performance in the occupation and are consistent with the purpose for which the licensing or certification program was instituted. (p. 13–2)

The comment following Standard 13.1 emphasizes the importance of job analyses and, in particular, of relating the knowledge and skill covered by the examination to the requirements of practice:

The fact that successful practitioners possess certain knowledge or skills is relevant but not persuasive. Such information needs to be coupled with an analysis of the purpose of the licensing or certification program and the reasons that the knowledge or skill is required for competent performance in the occupation. (p. 13–3)

As suggested by this comment, the purposes of licensure and certification are sufficiently different from those in employment testing and sufficiently important as to merit the development of job analysis procedures that are specifically tailored to the purposes of licensure and certification. As discussed later in this chapter, I would expect these specialized job analysis procedures to incorporate the results of previous research (e.g., clinical trials) and logical analysis of the demands of practice in addition to the kinds of empirical job descriptions usually derived from job analyses in employment settings.

The other standards in chapter 13 of the draft Standards emphasize disclosure policies and issues, like reliability and reading level, which are related to how well the examination measures the knowledges and skills included in the content domain for the examination. In general, the approach taken here is consistent with the fourth draft of the Standards, which is expected to be similar to the final form of the Standards.

Uniform Guidelines. A validation strategy based on critical abilities is also consistent with the Uniform Guidelines for Employee Selection Procedures (EEOC et al., 1978), which are used by the federal agencies in enforcing civil rights legislation. There is some question about whether the Uniform Guidelines apply to state licensure examinations. As stated in question 7 of the Adoption of Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures (Equal Employment Opportunity Commission, Office of Personnel Management, Department of Justice, Department of Labor, & Department of the Treasury, 1979):

7. Q. Do the Guidelines apply to the licensing and certification functions of state and local governments?
A. The Guidelines apply to such functions to the extent that they are covered by Federal law. Section 2B. The courts are divided on the issue of such coverage. The
Government has taken the position that at least some kinds of licensing and certification which deny some persons access to employment opportunity may be enjoined in an action brought pursuant to Section 707 of the Civil Rights Act of 1964 as amended. (p. 11997)

However, even if these guidelines are not legally binding, they are likely to be employed in legal reviews of testing procedures, and they have been made part of state law in California (Werner, 1981).

The Uniform Guidelines, which were developed primarily for employment selection, emphasize criterion validity but allow for procedures measuring specific abilities if it can be shown that:

(a) the selection procedure measures and is a representative sample of that knowledge, skill, or ability; and (b) that knowledge, skill, or ability is used in and is a necessary prerequisite to performance of critical or important work behavior(s). (p. 38302)

Therefore, the Uniform Guidelines explicitly allow for selection tests based on critical abilities, and as argued here, this approach is especially appropriate for licensure examinations.

**NCHCA Guidelines.** The National Commission for Health Certifying Agencies (1981) has published guidelines for credentialing examinations suggesting that certifying agencies should progress from content to predictive (or criterion-related) to construct validity.

This approach is laudable in setting ambitious goals for certifying agencies but may have some potentially negative consequences. In particular, by encouraging certifying agencies to take predictive validity and construct validity as goals, the NCHCA guidelines may draw attention away from the basic issue of content relevance. Since I am not optimistic about the value for licensure and certification examinations of predictive validity and versions of construct validity that require the adoption of strong theoretical assumptions, I think that this would generally be a bad trade-off if it occurred. The critical abilities approach to validation incorporates aspects of content, predictive, and construct validity, and aims to develop a validation strategy specifically designed for licensure and certification examinations.

**CHANGES NEEDED**

As is probably clear by now, the basic theme of this discussion is that the validation of licensure and certification examinations should be tailored to the purpose of these examinations and should be consistent with the intended in-
interpretations of the examinations. Given this assumption, it was argued in the last section that criterion validity is inappropriate for licensure and certification examinations for both practical and conceptual reasons, but that a strategy based on critical abilities is both feasible and consistent with the intended interpretations of such examinations.

A validation strategy based on critical abilities incorporates many elements of content validity. The standard method for establishing the content validity of tests is to have experienced practitioners decide which abilities need to be evaluated. These content decisions may be based in part on empirical studies of patterns in the conduct of practice in order to ensure that the content of the examination reflects the actual day-to-day practice of the profession. Both expert judgment based on the department of learning and empirical job analysis may play large roles in studies of content validity but do not supply all of the evidence needed to establish the validity of licensure examinations as measures of critical abilities. In addition, there are a number of issues, in particular, the relationship between the critical abilities and performance in practice, that an argument for the validity of a licensure or certification examination should address.

Abilities Should Be Clearly Related to Client Outcomes

The abilities measured by the examination should be "critical" in the sense that they have a significant influence on client outcomes, and any ability required for licensure should be explicitly linked to client outcomes. The linkage may be based on clinical research, on logical analysis, or on a combination of the two, but it should be explicit.

For many professions the linkage between critical abilities and client outcomes has a large empirical component. The requirement that pharmacists be able to dispense drugs correctly is based on clinical research relating dosage to the effectiveness and safety of the drugs. The expectation that physicians know the symptoms and typical course of development of various diseases is based on empirical research showing that the detection and subsequent treatment of the diseases has positive effects. In some professions, the linkage between various abilities and client outcomes is based mainly on logical analysis. A strong logical case can be made for the linkage between knowledge of the law and effectiveness in such professions as accounting or law. Similarly, the relationship between knowledge of academic content and the ability to teach that content is based more on logical analysis than on empirical studies. Generally, the critical abilities will be determined by a combination of empirical data and logical analysis that constitutes the department of learning for the profession.

Where evidence linking an ability to client outcomes is less straightforward than it is for these simple examples, decisions about criticality become more complicated. If there are several approaches to some issue of professional practice and the evidence does not consistently favor one approach, it would still be
reasonable to require that candidates for licensure know enough about the various approaches to recognize their potential benefits and limitations. Given that none of the approaches is clearly superior to all others, it is necessary to allow for the use of professional judgment in selecting a particular approach for each client, but it is also appropriate to require that practitioners be familiar with the available options. Given the purpose of licensure, it is especially important that practitioners be aware of any dangers inherent in various interventions.

Knowledge of research results and theory provides a basis for informed clinical judgment. Since the situations encountered in the practice of most professions tend to be highly variable, the most effective approach to each situation cannot be standardized, and the practitioner is often called upon to employ professional judgment. For many areas of practice, mastery of a domain of knowledge that is relevant to a broad range of situations may be required to inform the practitioners’ decisions about how to handle specific situations. That is, an approach to validity based on critical abilities should not be viewed as an attempt to reduce highly developed, complex content domains to a set of discrete pieces of knowledge linked to specific practice situations. The linkage of theory to client outcomes may be more general and less direct than it is for specific skills, but the linkage should be clearly established.

Abilities Should Be Weighted According to Their Importance for Practice

Since some critical abilities will be more important than others, the weight given to various content areas in a licensure examination should reflect the importance of the content areas for practice. The importance of an ability depends on how often it is needed in practice and on how much difference it is likely to make in terms of client outcomes.

The frequency of occurrence of a situation in professional practice is obviously one factor in determining how important it is that a practitioner be able to deal with the situation. For example, it is clearly appropriate that examinations for medical licensure in the United States devote considerable attention to heart disease, diabetes, cancer, and flu because they have a high rate of incidence. The content of examinations used to certify practitioners for specialized practice would naturally give a heavy emphasis to the conditions included in the specialty area even if these conditions are not encountered often in general practice. But even here, the more common conditions in the specialty would generally be given more emphasis than rare conditions.

Job analyses usually place heavy emphasis on frequency data (Williamson, 1979). There are several empirical methods for determining the frequencies of occurrence of various situations in practice. The most obvious method is to ask practitioners how they spend their working hours (e.g., see Williamson, 1979). A more direct approach involves observing the professional’s activities over an
extended period (Miller, 1963). The direct observation approach is less subject to
the kinds of bias that often occur in self-reported data but is more expensive to
implement, and it is therefore likely to involve a smaller, and perhaps less
representative, sample of practitioners. Both of these approaches provide data on
the kinds of demands placed on practitioners and on the time devoted to different
kinds of activities, and are therefore clearly relevant to the issue of content
validity.

In addition to formal job analyses, there are often existing sources of informa-
tion about the demands encountered in professional practice. For example, the
statistics routinely collected by local, state, and federal government provide a
wealth of information about the incidence of various health problems (e.g., a
morbidity and mortality weekly report is published by the Center for Disease
Control in Atlanta); these data indicate the kind of patients that are likely to be
encountered in the practice of the health professions, and therefore provide data
relevant to the frequency with which various situations will be encountered in
practice.

A major difficulty with data on how professionals spend their time is that the
activities included in such data will vary in their importance relative to the
purpose of licensure, protection of the public. Williamson (1979) reports that
32% of a physician’s time at work is spent on activities other than patient care.
Even if attention is restricted to the abilities required in providing professional
services to clients, frequency data do not indicate how serious the lack of an
ability would be in a particular situation.

The second component in evaluating the relative importance of different
abilities is the gravity of the possible consequences of the situations that require
the ability. Although common colds occur more frequently than concussions, the
consequences that would result from improper treatment of a concussion suggest
that a licensing examination for physicians should give more attention to the
concussions than the frequency of this condition might indicate. This is not to say
that the treatment of colds should be ignored, but rather that the weight given to
various abilities should be a function of both frequency and seriousness. Rakel
(1979) has stated this point succinctly:

The temptation to achieve content validity in examinations by matching test items
to the frequency of problems encountered in practice could also be counterproduc-
tive. There is a justifiable need to test more heavily on problems that have a high
morbidity and fall into the “uncommon but harmful if missed” category. Because
of their serious nature, they deserve greater representation in an examination than
practice surveys indicate. (p. 93)

Given that the purpose of licensure is to protect the public, the “harmful if
missed” category should be emphasized in licensure examinations. Licensure
examinations should emphasize the abilities required by situations involving the
“uncommon but harmful if missed” category, as indicated by the department of
learning, even though these abilities are likely to have relatively low frequency of occurrence in practice.

Empirical job analyses are useful in providing data on relative importance as well as frequency. In addition to estimating the frequencies with which various situations arise, respondents are usually asked to indicate the criticality of the actions taken in each type of situation. The critical incident technique (Flanagan, 1954) specifically addresses the perceived importance of an activity as well as its frequency. However, this technique, which is focused on critical incidents, does not provide a clear-cut definition of professional competence. A serious limitation in empirical job analyses is that they focus on what is currently going on in practice, but they do not provide a thorough analysis of what practice would need to be like to best serve the public interest.

In order to address this larger issue of the public interest, the results of job analyses need to be interpreted in terms of the department of learning for the profession. The department of learning will generally provide the best guidance on how the various situations that arise in practice should be handled. Taking an example that is close to home, it seems unarguable that examinations used for teacher certification should reflect the best current thinking on how tests and other assessment instruments should be used in making educational decisions, and should not rely solely on surveys of current practice. In general, empirical job analyses are particularly useful in providing information about the kinds of situations that will be encountered in practice, while the department of learning for the profession is a more reliable source of information about how these situations should be handled. Therefore, in weighting various critical abilities, both empirical job analysis and the department of learning have major roles to play.

Extraneous Factors Should not Unduly Affect Examination Scores

The interpretation of test scores as measures of critical abilities assumes that differences in scores are due to differences among candidates in their attainment of the critical abilities. Cronbach (1980b) points out the need “to establish that an achievement test contains no irrelevant difficulty, if we are to say that it measures command of certain subject matter” (p. 106). To ensure that tests of the critical abilities are measuring what they claim to measure, plausible alternative hypotheses should be investigated.

Some potential competing hypotheses are examined under the heading of reliability or generalizability (Brennan, 1983; Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Measures of stability address the competing hypothesis that observed scores are a function of the occasion on which the measurement is made. Measures of interrater reliability address the hypothesis that scores are largely determined by the observer rather than by the candidate’s performance.
Measures of internal homogeneity (e.g., KR-20, coefficient alpha) or parallel-forms reliability address the competing hypothesis that the choice of a particular set of test items strongly influences the outcome.

A potentially serious problem for many types of measurement procedures is the presence of response sets, or tendencies of some persons to respond in a stereotypical way. Affective traits, which are defined in terms of typical performance rather than the best possible performance, are especially subject to response sets. A pattern of “correct” answers to questions about ethical issues may reflect a response set favoring socially desirable responses rather than a commitment to ethical behavior. In research on affective traits, this problem has sometimes been handled by camouflaging what is being measured, but the use of this approach in a licensing examination would raise serious practical and legal problems (Levine, 1980). As noted earlier, the fact that candidates for licensure and certification have a vested interest in performing well makes it especially difficult to evaluate affective traits like conscientiousness or “good character.”

Standard test development procedures (e.g., see Ebel, 1972) are designed to minimize the chances that candidates who have the abilities being tested will get an item wrong or that candidates who do not have the abilities being tested will get the item right. In particular, many of the rules for developing objective tests are designed to minimize the influence of response sets. Poorly constructed tests are likely to give an unfair advantage to candidates who are skillful at taking tests (Sarnacki, 1979).

It is also important to ensure that the language used in the test does not constitute an artificial barrier to performance. Except for the use of technical vocabulary, the reading level of the examination should be kept sufficiently low so that anyone with the abilities required for practice will be able to read it. Similarly, the instructions for the examination should be as clear and simple as possible (especially when the instructions are unusual, as they often are for simulations).

It is important to avoid any extraneous factors that could cause minority or women candidates who had developed the critical abilities being tested to get items wrong. Although important content should not be omitted simply to eliminate differences between subgroups, the wording of items should be reviewed to avoid any source of bias (Schmeiser, 1982) that would be likely to interfere with the performance of subgroups within the population.

Licensure and Certification Examinations Should Cover as Wide a Range of the “Critical” Abilities as Is Feasible

Since licensure laws typically qualify the professional to practice in a broad range of settings and to deal with the full spectrum of problems that arise in these settings, the results of the licensure examination are likely to be interpreted as indicating candidates’ command of a correspondingly broad range of abilities.
Although it is usually not possible to test the relevant domain exhaustively, the examination content should provide reasonable coverage of the domain as a whole. The scope of the content domain for certification examinations is likely to be more specialized than it would be for a licensure examination, but it is still important to sample the domain adequately. The test specifications typically used to develop licensure and certification examinations are designed to ensure a representative sampling of content. Of course, if the interpretation given to the results of the examination is consistent with a narrowly defined domain, the resulting examination could be highly valid as a measure of knowledge of that domain, but it would probably not serve the purpose of licensure examinations, the protection of the public, very well.

The content of the licensing examination must also be consistent with the scope of practice specified in the legislation authorizing licensure for the profession. Although the laws governing the scope of professional practice may be stated in general terms that leave considerable latitude for interpretation, the legal definition of professional practice still limits the content for a licensure examination to the extent that it limits practice. It would be inappropriate for a licensure examination to require demonstration of a skill that is legally prohibited in practice.

The Cognitive Level of the Items Should be Appropriate

Although the definition quoted earlier refers to "knowledge of some department of learning or science," it is clear that to be safe and effective in practice, the professional must also be able to use this knowledge to solve problems. The professional must be able to apply elements from the department of learning or science to the situations that arise in practice.

If the questions in a licensure or certification examination require the application of knowledge to specific situations, the performance required of the candidate taking the examination is closer to that required in practice than would be the case if the examination involved simple recall of facts. To the extent that the performance required on the licensure examination is similar to the performance required in practice, inferences drawn from test performance to readiness for practice are more direct and therefore easier to justify.

The Level of Proficiency Required by the Examination Should not Be Higher Than That Required for Practice

For a licensure or certification examination that measures a selected set of critical abilities rather than all of the characteristics required for good practice, it is important that the standards on the examination not be set unreasonably high. Although some level of mastery of a critical ability may be necessary for practice, it is not always true that higher levels of the ability will lead to improved performance. Thus, for example, some skill in arithmetic is necessary in many
professions, but mastery of higher mathematics would probably not improve performance significantly in most professions.

In general, licensure examinations emphasize knowledge and the ability to apply knowledge because these skills can be measured accurately with written tests. Since cognitive skills are important for professional practice, this is not inappropriate, but the standards for these abilities should not be higher than the level of competence required for practice. If the standards for the cognitive abilities are artificially high, the licensing examination is likely to exclude many candidates who would make good practitioners.

Validity Should Be a Public Function

As noted earlier, to validate an interpretation for an examination is to produce convincing evidence that the interpretation is justified. Since licensure examinations serve a public function, the evidence for validity should be public. That is, the types of evidence suggested in this section should be available for review by the public that licensure procedures, including the licensure examination, are designed to protect. Although the argument for public disclosure is not as strong for certification examinations, a reasonable level of disclosure would also be desirable for these examinations since their effectiveness depends to a large extent on public confidence in the certification process.

Where feasible, the release of sample copies of the examination would serve a useful function in informing discussion and debate about licensure and certification. The periodic release of retired forms of the examinations would not generally have a significant impact on the quality of the examinations and would provide an opportunity for external review of examination content, format, and quality. Complete disclosure of all examinations is probably not justified in most cases because of the costs involved and because the additional benefits of complete disclosure compared to partial disclosure would be marginal.

Since the evidence for the validity of licensure examinations is generally available to interested outside reviewers and since sample copies, or at least sample items from the examinations, are available, I don’t see the disclosure issue as a major problem. However, one area in which additional information is probably needed is standard setting. This is especially true because a major criticism of licensure examinations is that the passing scores have been used to restrict entry to the professions in order to protect the interests of the professions (Friedman & Friedman, 1980).

RECENT DEVELOPMENT AND NEW AREAS OF EMPHASIS

There are at least three areas in which improvements in the methodology applied to licensing and certification examinations are needed and, I believe, possible. Two of these have already been touched upon—standard setting and domain specifications. The third area involves the possibility of expanding the scope of
critical abilities that are included in the examinations through computerized simulations of practice situations.

**Standard Setting**

There are two basic problems with current judgmental methods for standard setting. First, the different judgmental methods for setting standards tend to yield different passing scores, and there is no good basis for choosing among them. In addition, different groups of raters using the same method yield different results. Second, the judgmental standard setting methods do not provide a clear basis for interpreting the resulting passing score; rather, the reference populations that provide the basis for norms-based interpretations are simply replaced by a new reference population of raters. Recent developments in judgment-based standard setting (e.g., see Jaeger, 1982) would involve more thorough surveys of the opinions of different types of raters and could therefore probably improve the stability of the results across replications of the procedure, but would not help with the second problem.

**Interpretability of Standards.** The judgment-based standard setting procedures do not yield an interpretation of what the standards mean in terms of what passing candidates can do, because the results are not explicitly tied to item content. In the Angoff procedure, for example, expert judges are asked to consider the expected level of performance on each item (the probability of answering the item correctly) of hypothetical “minimally competent candidates.” The judges are instructed to assign a minimum passing level (MPL) to each item in terms of the probability that a minimally competent candidate could answer that item correctly. Since the cutoff score for the examination is simply the sum of the MPLs for the individual items, it will depend on the sample of items and on the sample of raters.

Unless a behavioral interpretation of the test scores is available, the results of the Angoff procedure do not indicate the kind of behavior that distinguishes passing candidates from failing candidates. Although individual raters undoubtedly use some performance criteria in setting the MPL for each item, (e.g., their individual experiences with persons they considered to be minimally competent), the judgment-based standard setting procedures do not provide a mechanism for making these performance criteria explicit. Therefore, the interpretation of the resulting passing score depends on the criteria for selecting judges, and the burden of interpretation falls on the new reference population, the population of raters.

This concern about the interpretability of test scores in terms of explicit behavioral criteria is not new. Ebel (1962) suggested two methods for obtaining what he called “content standard test scores” over 20 years ago. One of his two methods is similar to the approach suggested below. Nitko (1980) has described a number of ways in which test scores can be referenced to specific content, some of which could be applied to licensure and certification examinations.
Of course, if some copies of the examination are made public along with the corresponding passing score, a reviewer could infer a behavioral interpretation by evaluating the content and difficulty of the examination and comparing the perceived difficulty to the passing score. The reviewer might even use one of the judgmental standard setting procedures to obtain an independent estimate of the difficulty of the examination.

**Improving the Interpretability of Standards.** An alternative approach that would make the interpretive information more readily available would be to provide data about the differences in performance of passing candidates and failing candidates on a representative sample of items. For example, on a written examination interpretive data of this kind might indicate the proportion of passing candidates and the proportion of failing candidates who correctly answered a particular item. If the topic is important, the question addresses a significant aspect of the topic, and passing candidates answer it correctly more often than failing candidates, such data would indicate that the examination is performing as intended. If the results were reversed and failing candidates did as well or better on the item than passing candidates, the data would suggest that the examination is not working as it should. Therefore, in addition to generating concrete referents for the distinction between passing and failing candidates, this kind of analysis provides a check on the overall validity of the examination process (e.g., see Council of State Boards of Nursing, 1979, pp. 123–127.)

A somewhat more sophisticated approach would be to provide graphs of the proportion of candidates answering an item correctly as a function of total score on the examination. Such graphs would provide detailed information about the implications of total test score for performance on that particular item and would therefore say something about the consequences of setting the passing score at a particular level. Graphs of this kind for a representative sample of items would provide a basis for the interpretation of the test scores in terms of candidate performance on the skills tested by the items.

**A Check on Validity.** In addition to its impact on interpretability, such approaches could lead to improvements in the setting of standards by providing a check on the internal consistency of the ratings. The minimal pass level (MPL) for an item represents the probability that a “minimally competent examinee” would be able to answer the item correctly. The passing score for the test is the sum of the MPLs for the items in the test. According to the assumptions underlying this procedure, candidates with scores at or just above the passing score can be considered minimally competent. By examining the proportion of these candidates who answer a given item correctly, we obtain an estimate of the probability that a “minimally competent candidate,” as defined by the Angoff procedure for the test as a whole, can answer the item correctly. To the extent that this probability differs from the original MPL for the item, there is some inconsistency in the results.
Perfect consistency on judgments about MPLs for different items is not to be expected, and experience with the approach will be needed to determine what constitutes adequate agreement. However, major inconsistencies would suggest a possible problem; for example, if an item is important enough that the raters think that minimally competent candidates should be sure of answering it correctly (i.e., the MPL is 1.0), but candidates with relatively low probabilities of answering the item correctly are passing the examination, the passing score may be too low. At the very least, such comparisons will inform the raters of the fallibility of the standard setting procedure. It also gives the raters the opportunity to reexamine the overall passing score in light of its implications for particular items. This approach would be a natural extension of the Angoff procedure, which is based on raters’ judgments of the probability that a minimally competent candidate would get an item correct, but it could be used for any judgmental standard setting procedure.

Because this approach has not been tried out yet, I would not recommend it for immediate application. However, I do think that it would be a fruitful topic for further research.

Definition of Content Domains

The task of defining an appropriate content domain for licensure and certification examinations is extremely important, but the methodology for accomplishing this task is not highly developed. However, in part because I have already discussed it to some extent, and in part because I do not have a very definite program for improvement to propose, I will restrict myself to a few general remarks on this topic.

First, we need to face the fact that the definition of the content domain, like the setting of standards, involves judgments and is therefore subjective to some extent. Attempts to substitute data for judgment in defining the content domain may succeed in diffusing responsibility for the judgments, but it doesn’t necessarily improve the domain definition or the examinations developed to reflect the content domain.

I raise this issue as an important focus for study in part because of the emphasis that has sometimes been given to empirical job analyses as a necessary component of content validity (e.g., see Equal Employment Opportunity Commission et al., 1978). Data on how practitioners spend their working time are clearly relevant to the definition of content domains for licensure and certification examinations because they indicate the frequency with which various situations occur in practice. Most job analyses generally collect data on practitioners’ ratings of the importance as well as the frequency of various activities, and therefore provide information about practitioners’ perceptions of what aspects of current practice are most important. Therefore, empirical studies of patterns of practice provide valuable guidance in specifying the range of situations encountered in practice and can be supplemented by the extensive data available, for
some professions, at least, on the incidence and severity of various kinds of problems with which practitioners are expected to deal. Such studies indicate what is going on in practice, but unless we adopt the view that "all is for the best in this best of all possible worlds," they do not indicate what should be going on. Given that various activities are given different levels of emphasis in practice, the appropriateness of this distribution of emphasis is still open to question, and the resolution of such issues involves complex judgments.

The Role of the Department of Learning. Given specific situations, decisions about the knowledge and skills needed to deal effectively with three situations can be based on the relevant department of learning. The profession is defined in terms of the department of learning, which provides a body of theoretical and empirical knowledge of the causes, likely courses of development, and appropriate interventions for the situations encountered in professional practice. Such information provides a basis for identifying knowledge and skill required to deal effectively with situations resulting from the job analysis. For example, assuming that one determined, by an empirical job analysis or logical analysis, that a significant part of CPA practice involved the preparation of tax returns, it would probably be better in terms of validity to base a CPA examination on what the federal and state tax codes say can and should be done in determining tax liability than on surveys of what is done; the empirical job analyses would indicate which parts of the tax codes deserve most emphasis, but the items would be based on the tax code.

The most serious limitation in the use of the department of learning is that it is often unwieldy because it is extensive and is not organized in a way that is convenient for test development. The development of a test plan from the department of learning and a job analysis is usually accomplished by content specialists drawn from the profession. However, a readily available and organized source of information that can facilitate the translation of the department of learning into a domain definition for the examination is the textbooks used in professional schools. For reasons outlined earlier, licensure and certification examinations should not simply follow the curricula of professional schools, but the content of these curricula presumably reflect the combined judgments of faculty about what practitioners need to know. If professional school faculty are totally misguided about the demands of practice, society has a much more serious problem than the misspecification of the content domain for licensure and certification examinations. Therefore, textbooks may provide a useful source of information in defining the content domain for licensure examinations.

Empirical Check on Content Domain. A potentially useful, albeit expensive, procedure for empirically evaluating the procedures used to specify content domains was discussed by Cronbach (1971). Applying this general approach to licensure and certification examinations, two versions of the content domain and
resulting examinations could be developed independently, based on the common objective of evaluating professional competence, using similar but preferably not the same procedures. The detailed specifications of the content domain and corresponding examinations would be developed by different groups of content specialists and test development experts, and the scores on the two examinations for a sample of candidates would be compared. If the scores on the two examinations were in good agreement, we would have evidence that the choice of experts, the details of the content domain specifications, and the procedures used for examination development do not have an undue influence on the outcome.

Impact of Public Policy on Content Domains. In using information from a job analysis and department of learning to develop licensure examinations, it is important to keep in mind that licensure is a public function controlled by law. Although professional practitioners and content specialists necessarily have much to say about requirements for licensure, including examination content, the public and, more specifically, the public’s representatives in state legislatures also have a major interest in such requirements. A change in state law requiring that certain topics be taught in the high school science curriculum (e.g., content relevant to alcohol and drug abuse) would clearly have implications for teacher certification; such requirements, which are motivated by a desire to address social problems, would not necessarily be reflected in current job analyses or in content experts’ judgments of what constitutes the core of their academic discipline. The point of this example is not that such specific requirements are commonly included in licensure laws; they are not. However, licensure laws do provide the legal basis for licensure, and although such laws are stated broadly, they incorporate a general view of requirements and restrictions in the practice of the profession being licensed. If the licensure procedures are to follow legislative intent, the content domain for the examination should be consistent with this general view of professional practice.

Obviously these remarks do not constitute a model for content domain specification. At best, they reflect some issues that could be considered in developing such a model.

Expanding the Scope of the Content Domain
My last suggestion of an area for future investigation may be too obvious to mention, but I will do so anyway. The suggestion is that it would be desirable to expand the scope of what is assessed on licensure and certification examinations to give more emphasis to realistic applications of professional judgment and less emphasis to factual knowledge. Current technological developments may offer good opportunities to do so more efficiently and more effectively than has been possible in the past. In particular, computerized simulations of practice situations may provide an effective means for evaluating skills that are not easily assessed in printed examinations (see McGuire, Solomon, & Bashook, 1976).
Simulations generally begin with a description of a client and the circumstances under which the client is first encountered, followed by a series of questions about what actions the examinee would take in order to assist the client. After the examinee has chosen an action, feedback on the results of the action is provided. As the test progresses, the situation is developed by providing additional information in the questions and in the feedback that is provided to the examinee after each response. The aim is to make the descriptions of the situations as realistic as possible and to require that professional judgment be used in deciding what to do. An advantage of simulations is that they make it possible to observe "performance" for a large number of simulated clients within a reasonable period of time.

The technology available for use with simulations includes relatively inexpensive microcomputers that are capable of presenting simulated situations and monitoring candidates' performance as they attempt to deal with the problems presented. It also includes videodisc equipment which can present high resolution photographs, as well as video segments involving sound and motion. Therefore, this technology may make it possible to present standardized but highly realistic simulations on an individual basis.

Licensure and certification examinations tend to be quite long, and in many cases good estimates of candidates' mastery of the content domain could be obtained with fewer items. The efficiency of many of the examinations could probably be further enhanced by a judicious allocation of items to subcategories in the domain (see Jarjoura & Brennan, 1982). Therefore, by diverting some of the resources currently devoted to developing multiple-choice items, computerized simulations might be used to expand the range of skills included in the content domain without significantly lengthening the examinations.

Limitations of Simulations for Assessment. In the short term, the practical difficulties of having enough terminals for candidates, maintaining security, and developing software will limit the applicability of this approach. There are also some conceptual problems associated with the use of simulations in licensure and certification examinations that need to be addressed. For example, in branching simulations, a candidate who chooses an option that introduces complications would be asked to deal with these complications and therefore would be required to demonstrate skills that other candidates might not be required to demonstrate. This raises issues of comparability that would not arise in a multiple-choice examination or in a linear simulation, where everyone answers the same questions. Given the importance of both fairness and the appearance of fairness in licensure and certification examinations, this lack of comparability may be viewed as a problem.

A related issue is the adequacy of sampling of content. Because realistic simulations take a significant amount of time to work through, the examination is likely to involve a relatively small number of separate simulations, thus making
it difficult to assess a broad domain of content. This problem might be ade-
quately resolved by using multiple-choice items to achieve broad coverage of the
domain, while emphasizing the assessment of problem-solving strategies and
professional judgment in the simulations.

A potentially serious problem in using simulations as a major component of
licensure and certification examinations is their reliability. To the extent that the
simulations are highly realistic and involve a high degree of professional judg-
ment, assessments of candidate performance is likely to be variable across simu-
lations. The result would be low reliability reflecting the fact that each can-
date’s score would be determined to a substantial degree by the choice of
simulations employed rather than by the candidate’s overall competence. Since
the separate responses called for by each simulation are not independent (e.g., a
candidate who gets off on the wrong track might find it very difficult to achieve a
satisfactory performance on the simulation as a whole), the simulation examina-
tion would consist of a relatively small number of “items” (i.e., simulations).
Therefore, the improvement in reliability achieved by averaging across a large
number of independent items that is possible in multiple-choice examinations
would not operate for simulation-based examinations.

In spite of these potential difficulties, I think that simulations offer consider-
able promise for extending the range of abilities that can be reliably assessed in
licensure and certification examinations. It may be necessary to make some
compromises between realism and standardization of the content covered, but
further research on the properties of simulations should lead to improvements in
their effectiveness as assessment instruments.

**SUMMARY**

There are two basic themes that have guided my discussion of licensure and
certification examinations. The first of these themes is that the approach taken to
validating the examinations should reflect the intended purpose and interpreta-
tion of the examinations. The second theme is that the results of research on the
validity of the examinations and related issues, such as standard setting, should
provide a basis for public discussion of these issues and cannot be expected to
provide simple answers to complex questions.

**Validation Strategy**

Matching the validation approach to the intended interpretation of the examina-
tions requires an explicit description of the intended interpretation. I have argued
that the interpretation of the examination scores in terms of abilities that are
necessary for safe and effective practice is appropriate, given the purpose of
licensure, the protection of the public.
Given that an interpretation in terms of critical abilities is adopted, the argument for validity involves two components. The first component requires evidence that the abilities to be covered by the examination are critical for practice. Empirical job analyses are particularly useful in identifying situations that commonly arise in practice and in providing some data about the consequences of proper and improper handling of such situations. The department of learning provides information on how these situations should be handled to obtain optimal results. In combining these two sources of data, professional judgment plays a large role.

The second component requires evidence that the test scores reflect competence in the domain of critical abilities defined in the first step. Much of the evidence relevant to this issue is derived from the critical examination of the properties of the test in the light of plausible counterhypotheses about what the test measures.

As noted earlier, the critical abilities approach involves aspects of content validity, criterion validity, and construct validity. Each of these three standard validation strategies focuses on a specific type of interpretation, and the interpretation of licensure and certification examinations in terms of critical abilities involves aspects of all three of these standard interpretations. The methodology of content validity is central to defining the content domain and providing evidence that the examination samples the domain adequately. The emphasis on establishing the link between the critical abilities included in the domain and practice involves indirect criterion validation. The examination of plausible counterhypotheses involves issues usually considered under construct validity.

Informing Public Discussion

The second major theme of this discussion is that licensure and certification are public functions, subject to public scrutiny. The appropriate scope of practice for licensed professionals is a matter of public policy, which is encoded, at least in general terms, in licensure laws. The scope of practice reflected in voluntary certification procedures is determined by the professional judgment of the certifying agency but is also shaped by public expectations.

Therefore, research on the validity of licensure and certification examinations is more akin to policy research than it is to basic scientific research. The situations that practitioners should deal with, the types of interventions that they should employ, and the standard of skill to be expected in the implementation of these interventions are not questions that can be answered by empirical data; they are policy issues.

This suggests that the results of validity studies that attempt simply to determine whether the test is valid will be less helpful ultimately than studies that provide information useful in making policy decisions and, more fundamentally,
in informing debate about policy issues. Questions particularly relevant to policy decisions include:

1. What types of situations occur most frequently in practice and/or have the most serious potential consequences?
2. What abilities, including knowledge and skills, are needed to deal with these situations effectively?
3. How well are these abilities taught and assessed in the educational programs preparing practitioners?
4. How well does the examination assess competence in the critical abilities, and what sources of variance other than differences in competence (e.g., reading level, response sets, ethnic bias) influence examination scores?
5. What are the implications of setting standards at different levels?

Although research on the validity of licensure and certification examinations will not resolve such issues, by addressing questions like those listed above, it can promote informed decision making.

REFERENCES


