DOCUMENT RESUME

ED 343 958 TM 018 187

AUTHOR Kane, Michael T.
TITLE The Validity of Assessments of Professional Competence.
PUB DATE 92
NOTE 30p.
PUB TYPE Viewpoints (Opinion/Position Papers, Essays, etc.) (120)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Competence; *Objective Tests; Observation; *Performance Tests; *Personnel Evaluation; *Professional Personnel; Rating Scales; Research Methodology; Simulation; Testing Problems; *Test Validity

ABSTRACT

Valid assessment of professional competence has proven to be an elusive goal. Objective tests, direct observation of performance, overall ratings of competence, and simulations have been tried and found wanting in one way or another. Objective test items are criticized as being unrealistic and therefore invalid. Direct observation tends to be very unreliable and thus invalid. Simulations and overall ratings of competence share both of these flaws to some extent. The difficulties inherent in evaluating professional competence are outlined, and some ways to minimize the impact of these difficulties are suggested. A general framework is proposed for evaluating the validity of measures of competence, and this framework is used to examine the strengths and weaknesses of three approaches to the assessment of professional competence: (1) direct observation; (2) simulation; and (3) objective tasks. In evaluating the validity of such an assessment, it is important to give special attention to the weakest links in the argument. For performance testing, evaluation and generalization are the weakest links. For objective tests, extrapolation is the weakest link. For simulations, any of the links can be weak or strong depending on the simulation. There is a 28-item list of references. (Author/SLD)

The Validity of Assessments of Professional Competence

Michael T. Kane

Abstract

Valid assessment of professional competence has proven to be an elusive goal. Objective tests, direct observation of performance, overall ratings of competence, and simulations have been tried and found wanting in one way or another. Objective test items are criticized as being unrealistic and therefore invalid. Direct observation tends to be very unreliable and therefore invalid. Simulations and overall ratings of competence share both of these flaws to some extent. Basically, you can't win.

This paper outlines some of the many ways to lose, and some ways to cut these losses. In doing so, it proposes a general framework for evaluating the validity of measures of competence, and uses this framework to examine the strengths and weaknesses of three approaches to the assessment of professional competence: direct observation, simulation, and objective tasks.

The assessment of professional competence is a difficult and, in many ways, a very frustrating task. It is difficult because professional practice is a complex and intellectually demanding activity, and is, as a result, not easy to describe very precisely or to evaluate accurately. Experts have been known to disagree about how to handle specific situations that arise in professional practice, making it difficult to evaluate an examinee's performance in that situation. This inherent difficulty is exacerbated by the impact of client/situation variables on professional performance, because the variability in performance across clients and situations makes it difficult to draw accurate conclusions about a practitioner's general level of competence based on a sample of performance.

Yet, on the face of it, the assessment of professional competence does not seem very difficult. We all have a general sense of what we mean by competence in various endeavors. Experienced professionals have a good, general understanding of the demands of professional practice and can identify clear instances of both good practice and bad practice. So, competence assessment looks easy! Difficult tasks that look easy tend to be frustrating.

Unfortunately, we cannot eliminate the sources of this frustration. The best we can do is to clarify the difficulties inherent in evaluating professional competence and, perhaps, suggest ways to minimize the impact of these difficulties. It has been said that the purpose of inquiry is <u>not</u> to move from confusion to understanding, but to move from confusion to a higher state of confusion, one in which we clearly understand what we are confused about. In this vein, my purpose here is to examine, from a psychometric point of view, the inferences involved in assessing professional competence and the sources of error that can undermine these inferences. This will not eliminate

the difficulties, but it may help to clarify why the assessment of professional competence is difficult and may make efforts to control errors of measurement more effective. My aim is to move to a higher state of confusion.

In the next section, professional competence is defined as the ability to use professional knowledge and skills to solve the problems that arise in practice. In the following section, I propose a general framework for the validity of measures of professional competence. This model treats validation as the evaluation of the inferences drawn from test scores and on the three major inferences involved in going from assessment scores to conclusions about competence: evaluation, generalization, and extrapolation.

The model assessment for the validity of professional competence assessments is then used to examine the strengths and weaknesses of three commonly used assessment methods: observations of performance, simulations, and objective tests. As one might expect, none of the three methods gets a perfect report card; they all have strengths and weaknesses.

Professional Competence

What do we mean by "competence" in a profession? This may seem to be a superfluous question; the word "competence" is commonly used in many contexts and is not considered especially obscure. In fact, most discussions of competence assessment in the professions assume that the term "competence" does not require explication. However, some discussion of the nature of "competence" is necessary in order to be clear about what it is that we want to assess.

The level of an individual's competence in some area of practice can be defined in terms of the extent to which the individual can handle the various situations that arise in that area of practice. Using the terminology of

LaDuca, Taylor, & Hill (1984) I shall refer to such situations as "professional encounters" or "encounters", where each encounter involves a context, a client, and the reason (the goal or problem) for professional intervention. Professional encounters vary in terms of the problem to be addressed, in terms of client characteristics (e.g., age, sex, level of functioning) and, in terms of context/setting variables (e.g., availability of resources and support personnel).

In defining an area of practice, it may be useful to give greater or less emphasis to different kinds of encounters depending, perhaps, on their frequency of occurrence or their degree of importance/criticality, or on both of these factors. Nevertheless, the area of practice can be described in terms of a domain of professional encounters, and to be competent in the area is to be able to handle the encounters in this domain.

This is a very fundamental way to think about competence (McGaghie, 1991; LaDuca, Engel, & Risley, 1978). Clients have needs for professional help; the purpose of the profession is to provide such help; and practitioners are competent to the extent that they can provide appropriate help to the client. For example, according to Norman (1985) "the family physician is competent to the extent that he can manage the problems he is likely to encounter: management of emotional problems, the problems of detection and compliance in the 10 percent of adult practice who are hypertensive, and so forth" (p. 25).

However, to say that professionals are competent is also to say something that goes beyond their expected performance over some domain of encounters. Competent professionals are expected to help clients by using certain professional tools, including subject matter knowledge, procedural

knowledge and skills, and the judgment needed to combine various knowledges, skills, and abilities into effective solutions to client problems (Benner, 1984). McGaghie (1980) treats "competence" as a semantic label that, "...connotes knowledge, skill, and acumen. It is assumed to be a general attribute of high-ability professionals" (p. 295).

The knowledge base of a profession is typically well developed and highly sophisticated, and often has a long and illustrious history. The organization of curricula in professional schools both reflects the organization of the knowledge base and tends to institutionalize it. For all of these reasons, the knowledge base shapes our thinking about professional practice and professional competence.

There are, then, two components in conceptions of competence. One component is the domain of possible encounters that the professional is expected to manage effectively, and the other component includes the knowledge, skills, and judgment that the professional is expected to use in managing these encounters. Combining these two components, an individual's level of competence in an area of practice can be defined as the degree to which the individual can use the knowledge, skills, and judgment associated with the profession to perform effectively in the domain of possible encounters defining the scope of professional practice.

This definition will provide us with a target variable, or a "conceptual criterion", in analyzing the potential weaknesses and strengths of various methods for evaluating professional competence. The target is "conceptual" in the sense that it is not possible to measure competence, as defined here, directly, and therefore this criterion cannot be used to empirically validate

measures of competence. Rather, the definition provides a basis for examining the inherent limitations of different approaches to measuring competence.

The Validity of Competence Assessments

Validity is the primary concern in evaluating assessment procedures. According to the Standards for Educational and Psychological Testing, (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1985) validity "...refers to the appropriateness, meaningfulness and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences" (p. 9).

The purpose of an assessment of professional competence is to provide an indication of an examinee's ability to use the appropriate knowledge, skills, and judgment to provide effective professional services over the domain of encounters defining the area of practice. If we could observe each examinee's performance over the full range of encounters in the domain of professional encounters and could evaluate the performance in each encounter unambiguously, the interpretation of the results in terms of competence would be valid by definition. However, it is generally impossible to implement this type of exhaustive assessment. The scores generated by actual assessments are based on a limited, and perhaps nonrepresentative sample of performance, observed under conditions that may or may not be similar to those commonly found in practice. The evaluation of each performance is based on judgments about its appropriateness and effectiveness. The validity of an assessment of professional competence depends on the evidence supporting inferences from an examinee's score, which is based on fallible evaluations of limited samples of

performance, to conclusions about the examinee's expected performance over the domain of encounters defining the area of practice.

If definitive assessments of competence were available, the validity of any other assessment procedure could be evaluated by comparing its results to scores on the definitive assessment, thus generating criterion-related validity evidence, with the definitive assessments constituting the criterion measure. However, because such definitive assessments are not available, this approach is not feasible. As an alternative, the validity of the interpretation can be examined by evaluating the plausibility of the chain of inferences involved in going from the assessment scores to conclusions about competence. Taking this approach, our general definition of competence serves as a conceptual "criterion" in evaluating the plausibility of the interpretation.

The plausibility of an interpretation depends on the possible weaknesses in the interpretation and on the availability of evidence supporting the interpretation (Cronbach, 1971; Messick, 1989). As Cronbach (1980) has said, "The job of validation is not to support an interpretation, but to find out what might be wrong with it. A proposition deserves some degree of trust only when it has survived serious attempts to falsify it" (p. 103). Therefore, in evaluating the validity of assessments of professional competence, it is important to look for possible flaws in the chain of inferences from the results of the measurement procedure to conclusions about competence. In the next section, we shall develop a framework for evaluating the validity of assessments of professional competence in terms of the inferences involved in drawing conclusions about professional competence based on the assessment results.

Interpreting Scores as Measures of Competence

The data used as the basis for assessment generally involve observations of the professional's responses to certain problems. The problems could be straightforward questions about the most likely conclusions to be drawn from certain data, or they could be actual client problems that need to be analyzed and managed. The problems could be presented as multiple-choice items, as simulations, or as actual practice situations. But, in each case, there is some "problem" to be solved.

The examinee responds to the problem in some way. For a multiple choice question, the examinee would respond by recording an answer on an answer sheet. For direct observations of performance in practice and for some kinds of simulations, the examinee could respond by actually doing something for a client (or simulated client).

The examinee's response is scored according to some rule or procedure. In the case of objective test questions, the scoring may involve a simple yes/no decision about the correctness of the response, based on a comparison with an official answer key. More generally, scoring involves judgments about the quality of the response (e.g., did the examinee identify all of the client's problems and deal with them appropriately?). Finally, the scores for the problems included in the assessment are combined in some way into a total score for the assessment (or into a series of subscores).

Interpreting an individual's score on an assessment procedure as a reflection of the individual's degree of professional competence requires at least three inferences: evaluation, generalization, and extrapolation. The first inference is evaluation of the performance at hand; deciding whether the observed performance is good, bad, or indifferent. The second inference is

generalization of the results from the observed performance to a universe of similar observations, for example, drawing general conclusions about skill in delivering babies from observations of one or two deliveries, or estimating expected performance over some domain of test questions from performance on the specific sample of questions in a test. The third inference is extrapolation of the results from the assessment context, which is always artificial to some extent, to conclusions about expected performance in actual practice. A serious flaw in any one of these three inferences can invalidate the interpretation as a whole.

Evaluation. The scoring rules used to evaluate performance necessarily embody some criteria for judging the quality of the examinee's responses. For objective-test questions, the criteria will focus on the correctness of the examinee's response, where the criteria for deciding on the correct response are based on the knowledge base of the profession. For observations of performance in practice, the criteria are likely to focus on effectiveness and efficiency in solving the client's problem and on the avoidance of any harm to the client. The specific criteria employed in evaluating performance will depend on many factors, including type of problems included in the assessment, the format of the assessment (written tests, simulations, performance tests), and purposes of the assessment (educational assessment, licensure, certification), but all such criteria need to provide a clear and credible basis for differentiating good performance from bad performance if the scores are to be interpreted in terms of professional competence.

Generalization. General statements about expected performance over some universe of observations, as distinct from factual descriptions or ratings of a specific observation, require inferences from the sample of observations to

conclusions about a larger universe of similar possible observations. Such inferences involve statistical generalizations based on sampling assumptions. So, for example, in using ratings of performance in actual practice situations as a measure of competence, the score based on the sample of performance ratings is generalized to the larger group of qualified raters, from which the raters actually used were sampled, and to the larger domain of possible encounters, from which the encounters actually used were sampled. The domain from which encounters were sampled for the assessment may be the domain defining the area of practice or it may be a subdomain; for example, the area of practice may involve the treatment of patients in hospitals, clinics, and patients' homes, but assessment may be restricted to the clinic. Similarly, objective-test scores based on a sample of questions are generalized to the expected performance over some universe of possible questions dealing with the same general content area.

Generalizations from a sample of observations to the universe of possible observations from which the sample is drawn are evaluated in terms of reliability or generalizability (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Brennan, 1983). The generalizability of assessment results can be estimated by examining the variation among independent replications of the measurement procedure on the same persons. Inconsistencies from one instance of the measurement procedure to another on the same person are attributed to errors of measurement. To the extent that the errors of measurement are large, inferences from a sample of observations to the universe of interest are undependable.

Extrapolation. The third inference is extrapolation from the behavior actually observed (e.g., performance on a written test or in a simulation) to

the behavior of ultimate interest (e.g., performance with actual clients in a real practice setting). Extrapolations from one kind of performance to a different kind of performance tend to be most plausible when the observed performance is very similar to the performance about which conclusions are being drawn, as when performance testing or highly realistic simulations are used to evaluate competence in professional practice. Extrapolation may also be based on claims that the skills included in the assessment are critical for successful performance in practice (Kane, 1986).

Note that all assessment involves some extrapolation. Even if performance in providing care to actual clients is directly observed, the fact that the performance is being observed may influence the performance. The observed performance may be better than the practitioner's typical performance if being observed causes the practitioner to be more thorough and careful than usual, or it may be worse if being observed makes the practitioner nervous. Therefore, even if the assessment involves the observation of actual professional performance, there is an extrapolation from the testing situation to non-testing situations.

The Characteristics of Three Assessment Methods

The discussion so far has developed a framework for thinking about the validity of assessments of professional competence. The framework was defined in terms of three types of inferences involved in interpreting assessment scores in terms of professional competence: evaluation of specific performances, generalization of these observations to a universe of possible observations, and extrapolation from this universe to the universe of encounters defining professional competence.

In this section, the framework is used as the basis for examining the validity of inferences about professional competency based on (1) direct observation of performance, (2) simulations, and (3) objective tests. The discussion of each of these assessment methods is quite general; the purpose is not to draw conclusions about the validity of any specific method, but to identify the potential strengths and weaknesses in each method. As we shall see, none of the methods is above suspicion.

(1) Direct Observation of Performance

Assessment of professional performance with real clients in real situations provides the most direct approach to assessing professional competence. Performance is rated by trained observers, generally using checklists or rating forms.

Evaluation. The assignment of scores to performances in real practice settings involves some serious problems. In dealing with complex practice situations rather than intentionally simplified "textbook cases", the best approach to a problem may not be immediately apparent, and the experts may disagree on the relative merits of different courses of action, thus making it difficult to grade performances in terms of their quality. The particular virtue of using the observation of performance in actual practice situations to assess professional competence is that it allows for the evaluation of the examinees' ability to deal with complex, realistic situations, but such situations pose the greatest difficulties in evaluating performance. These difficulties are reflected in the relatively poor levels of agreement among experienced raters often found for performance examination (Hubbard, 1971; Hoffman, 1977).

Because the criteria for evaluating performance in practice must apply to the wide range of situations that may arise in actual practice, these criteria are necessarily general and, therefore, considerable judgment is involved in their application. The subjectivity and potential for bias may be partially controlled by using checklists or rating scales and by training raters, but this subjectivity cannot be eliminated.

Generalization. In interpreting scores on performance tests as measures of competence, we generalize from observed performance on a sample of encounters to performance on a domain of encounters. Unfortunately, observing performance in actual practice settings is sufficiently inconvenient and expensive that the samples of performance are usually very small. Data collection typically occurs over a limited period of time, during which the examinee works with a limited number of clients in a specific context and is observed by one or two raters.

In situations involving real clients, real problems, and real situations, where control is limited and criteria for evaluating performance are necessarily quite general, ratings of performance will contain substantial errors of measurement. Independent replications of a procedure involving ratings of performance in practice might involve different types of professional encounters, different settings, and different raters. Since it is the general level of performance of the professional that is the focus of the assessment, differences that are observed from one professional encounter to another, from one setting to another, or from one rater to another would all be classified as errors of measurement. Because the variability in performance from one observation to another tends to be large and the number of observations tends to be small, inferences from observed performance over a

small sample of encounters to average performance over the domain of encounters defining competence may not be very accurate (Swanson, 1990).

Furthermore, the representativeness of the sample of encounters may be a problem, because the logistics of performance testing often dictate the choice of encounters (i.e., those possible in a particular setting on a given day), and these encounters may not be representative of the domain defining competence. Several separate client encounters may be included in the performance assessment, but these encounters are likely to be drawn from a single setting, and therefore the clients may have much in common. So, generalizing from a small, and possibly unrepresentative, sample of encounters to a broad universe of encounters represents a serious threat to validity.

Extrapolation. To the extent that the assessment is based on observations of performance in actual practice, and the testing process can be assumed not to alter performance, extrapolation is not a serious problem. However, the process of observing performance in and of itself may have some subtle and not so subtle influences on the quality of the performance. Some individuals may perform better in the testing situation than they would ordinarily, if, for example, being observed encourages them to be more conscientious or considerate of the client than they ordinarily would be. Others may do worse because they become self-conscious when they are observed. The examinee's awareness of being observed is likely to have an especially great impact if the testing situation involves the ethical component of practice; an examinee is not likely to neglect a client, divert drugs, or embezzle funds while being observed by one or more raters.

Overall Inference. For inferences about competence based on observations of performance in actual practice situations, extrapolation is

usually the strongest link. Evaluation can be a problem, and generalization is almost always a problem.

The distinction between evaluating a particular performance and the more general task of evaluating performance over the domain of encounters defining competence is important and easy to forget. There is a natural tendency to jump to conclusions based on small samples of performance, especially if the performances are particularly good or bad in some cases. This distinction accounts, in large part, for the fact that although experts can recognize good and bad examples of professional practice when they see it, it is difficult to develop an accurate measure of competence based on observation of performance. We can observe and evaluate a particular action (or performance in a particular encounter) directly, but performance in any particular encounter may be influenced by client characteristics and context variables that are beyond the control of both the evaluator and the professional practitioner. Because performance cannot be observed over the whole domain defining competence, inferences must be drawn from a very limited sample of observations to the domain of encounters.

As noted earlier, the reliability of scoring can be improved by having several raters score each performance and by using detailed scoring criteria. We can increase confidence in the generalization inference by using samples of observations that are as large and representative as possible.

We can improve the situation further by standardizing the encounters to some extent. We can, for instance, try to arrange things so that all examinees are observed working with a specific mix of clients--one client with one kind of problem, two clients with another kind of problem, etc. This kind of systematic sampling may improve the representativeness of the sample. It

also makes it possible to develop specific scoring criteria for each type of encounter included in the assessment and to train raters to use these criteria with some degree of consistency. Note, however, that even modest levels of standardization are difficult to implement in real practice situations, and these standardization efforts may also tend to make the performance assessment somewhat artificial and contrived.

Because evaluation and, especially, generalization are the weak links in the use of performance testing to evaluate professional competence, these two types of inferences deserve special attention in evaluating the validity of competence assessments that rely on this approach.

(2) Simulations

Tests based on simulations represent an effort to overcome the disadvantages of performance tests, while maintaining a high degree of realism or fidelity. 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 possible actions for managing the client's problem. After an action has been chosen, feedback on the results of the action is provided. As the test progresses, each choice elicits further feedback and leads to new choices among possible actions.

Some of the most common methods used to simulate professional encounters are: written patient management problems (McGuire, Solomon & Bashook, 1976); computer-based simulations, often used with videodisc (Melnick, 1990; Swanson, Norcini, & Grosso, 1987); and standardized patients (Stillman & Gillers, 1986; Stillman & Swanson, 1987). The aim is to make the simulated encounters as realistic as possible and to require that professional judgment be used in deciding what to do (Hubbard, 1978; McGuire, Solomon & Bashook, 1976).

Simulations are designed to be as realistic as possible, but practical constraints limit the degree of fidelity that can be achieved with any methodology. In addition, some aspects of the encounter may be purposely unrealistic in order to improve reliability. For example, the individual taking the simulation-based test may be prevented from taking certain actions because these activities would use up a lot of testing time but not provide much information about competence.

Evaluation. Because the simulation involves a well-defined problem, specific and detailed scoring criteria can be developed for each simulation, and raters can be trained in the use of these criteria. In addition, simulations can be designed so that extraneous factors (e.g., the availability of a particular piece of equipment in a particular setting) do not introduce errors of measurement. Both the client problem and the context can be standardized to a high degree.

Nevertheless, scoring problems are not unusual in simulations. To the extent that a simulation is a realistic portrayal of the complexities of practice, involving the impact of client history and preferences, context effects, resource limitations, time constraints, etc., the optimal solution for the problem is likely to be unclear. The more realistic the simulation, the harder it is to get the experts to agree in rating a performance. Swanson (1990) points out that:

> For a typical case, a broad range of patient management strategies are possible. Even if the simulation response to each of them is appropriate, it is difficult to develop scoring algorithms that appropriately reward alternative strategies that are equivalent in quality. It is also difficult to ensure that similar strategies differing in quality receive appropriate scores (p. 5).

In general, questions of style complicate the scoring of simulations just as they do for observations of actual performance, and scoring rules for simulations have tended to favor thoroughness over efficiency (Swanson, 1990).

The availability of explicit rules for assigning scores in each simulation tends to make scoring more objective for simulations than for performance tests, but the appropriateness of these rules may still be called into question.

Generalization. Observations of simulated encounters, like observations based on real encounters, tend to suffer from high variability from one case to the next (Elstein, Shulman, & Sprafka, 1978). However, generalization tends to be much less problematic for simulations than it is for direct observation of performance, because it is possible to evaluate examinees over a much larger, and potentially more representative set of encounters with simulations than would be possible in performance testing. Therefore, inferences from a sample of simulated encounters to the domain of simulations from which the sample is drawn tend to be more dependable for simulations than they are for performance tests. Nevertheless, generalization is a problem for simulation-based assessment unless a large number of simulations are included in the assessment.

Extrapolation. By definition, simulations do not involve real clients in real situations. If the simulation appears realistic, we may have confidence that performance on the simulation provides an accurate indicator of what the examinee would do in a similar situation in actual practice. However, even for high fidelity simulations the inference from a score to a conclusion about competence in practice is based on assumptions that are subject to doubt.

The empirical evidence supporting the relationship between performance on written simulations and performance in practice is not very encouraging (Feightner, 1985). Goran, Williamson, & Gonnella (1973) compared scores on patient management problems (PMPs) with the results of chart audits and report that performance on the PMPs was more thorough than actual performance as evaluated by chart audits and was not highly correlated with the chart-audit results. Page & Fielding (1980) report that pharmacists did a more thorough job of eliciting relevant information on a PMP than they did with simulated clients (who appeared without being identified as actors) in the pharmacist's regular workplace. Feightner (1985) expresses a serious concern about extrapolating PMP results to actual practice: "It seems clear that while PMPs simulate and approximate the clinical encounter, one cannot be certain that they allow a valid measure of an individual's performance in an actual clinical settings. Perhaps too much has been expected" (p. 195).

Overall Inference. Basically, simulations represent an attempt to have it both ways. An ideal simulation would be realistic enough to tap essentially the same knowledge, skills, and judgment required in actual practice, but would be standardized to promote reliability of scoring. Each simulation would be short enough so that many separate simulations could be used, thus providing adequate content coverage and supporting accurate generalizations. These goals are not necessarily incompatible, but their achievement requires a judicious choice of tradeoffs. A high degree of fidelity generally requires a fairly long simulation, and this limits the size and breadth of coverage of the sample of content involved.

Compared to performance tests, simulations tend to have larger and more representative samples of encounters and more objective scoring, but these

gains are made at the cost of making the encounters somewhat artificial. The first two inferences, evaluation and generalization, are strengthened, but the last inference, extrapolation, is more problematic. Depending on the quality of the procedures used, we may have more confidence or we may have less confidence in simulation scores as indicators of competence than we would in scores on a performance test.

(3) Objective Tests

Written tests are commonly used in licensure and certification to assess some aspects of competence (Shimberg, 1981). The scores on these written tests may be interpreted as measures of competence, or they may be given a more limited interpretation as measures of knowledge and skills considered critical for competence in practice (Kane, 1982; 1986).

Evaluation. Examinee responses on an objective test of knowledge and skills can be graded objectively, thus eliminating problems of subjectivity in scoring. The scoring keys are typically developed by panels of experts and reviewed by other panels of experts. The occasional standardized-test item that is scored incorrectly makes the news because we expect the scoring on objective tests to be perfectly accurate. It is an especially good story if a student finds a flaw after the item has been reviewed by several experts.

In practice, if any substantial disagreements arise about the scoring of an item on an objective test, the item is not likely to be used. If such disagreements arise after the test has been given, the item will probably not be included in the scoring of the test. These policies make the scoring keys for objective tests resistant to challenge. However, they may also make it more likely that the test will focus on factual questions or straightforward applications of well-established principles or procedures. It is difficult to

develop questions involving judgments about complex issues that will meet stringent criteria for objectivity, and therefore such questions may not get enough attention unless a very explicit attempt is made to include them (Swanson, 1990).

Generalization. Inferences from performances on a sample of objective items to some universe of such items tend to be highly dependable. It is possible for examinees to respond to several hundred objective items about different areas of content or different types of professional encounters in a few hours of testing. Because the precision of estimates of expected performance over the domain of items tends to be directly related to the number of questions in the test, objective tests, which may include several hundred independent items, can generate very precise estimates. Inferences from the sample to the larger domain are strengthened further by the fact that the content of each objective item can be selected independently and therefore objective tests can be designed to sample a wide range of content.

Psychometric theories are based to a large extent on statistical models, and it should not be especially surprising that the objective tests, which have been developed in conjunction with these theories, have excellent statistical characteristics. Objective tests are designed to facilitate generalization and, as a result, this step in the overall inference from test scores to conclusions about competence tends to be highly dependable. By contrast, the scoring on simulation tests and performance tests does not generate large numbers of independent responses; fewer responses are made, and within each simulation or observed client interaction, the responses depend on each other in complicated ways and therefore cannot generally be considered independent.

Extrapolation. Taking a written, objective test differs substantially from professional practice. McGaghie (1991) has made the point that in observational studies of medical practice:

> ...not one of the physicians studied ever has to answer complicated batteries of multiple-choice questions as a routine part of his or her professional practice, even though their competence evaluations are composed almost entirely of such items (p. 6).

Written tests may provide direct measures of certain enabling skills (knowledge and skills related to performance in practice) but, at best, they provide indirect indicators of what an examinee would do in a real practice situation.

The argument for extrapolation from scores on written tests generally claims that the cognitive skills and knowledge measured by the test are necessary (although probably not sufficient) for effective performance (Shimberg, 1981; Kane, 1982). In interpreting scores on the test as indicators of competence, we assume that the knowledge and skills measured by the test constitute an important subset of the knowledge and skills needed for competence in practice. Evidence for the relationship between the knowledge and skills measured on the test and the requirements of practice may be based on expert opinion or on the results of an empirical analysis of practice requirements, but the leap from performance on a standardized, written test to expected performance in practice is an inherently risky inference.

In some cases, it may also be possible to develop empirical evidence for the relationship between scores on the written test and some other measure of competence (e.g., a thorough performance assessment). However, there are severe practical problems in implementing this kind of study in a satisfactory

way (Shimberg, 1981; Kane, 1982) and, as noted earlier, the validity of the performance test as a measure of competence is open to challenge.

Overall Inference. Assumptions about the consistency and objectivity of scoring, and about the generalizability of the results, are all easier to justify for objective tests than they are for most other methods. The popularity of objective tests in large-scale testing programs is due, in large part, to the fact that they can be scored objectively and that the results can be generalized across test forms with a high degree of precision.

The weak link in drawing inferences from objective test scores to conclusions about competence is extrapolation from performance in answering a series of written questions to conclusions about actual performance in practice. In the absence of empirical evidence relating the test scores to a demonstrably valid criterion of performance in practice (which is hardly ever available), this inference must rely on assumptions about the similarity of the performance elicited by the objective items and performance in practice (which is fairly shaky in most cases) or arguments to the effect that the knowledge, skills, and judgments required by the objective test items are necessary, although probably not sufficient, for successful performance.

In evaluating the validity of objective tests as measures of professional competence, the main challenge is to establish a link between scores on the test and performance in practice. Because empirical studies of the relationship between test performance and a clearly valid measure of competence (i.e., criterion-related studies) are generally not feasible, the linkage will generally be indirect (Shimberg, 1981; Kane, 1982; Kremer, 1991), but the nature and strength of this linkage is the key question in validating objective tests as indicators of professional competence. By contrast,

collecting extensive data on the reliability of objective tests contributes little to the overall argument for validity because generalization over items, scorers, etc. is not likely to be a serious problem for objective tests; evidence for reliability is necessary, but it is not especially helpful to devote extensive resources to this issue.

Conclusions

The choice of method for evaluating professional competence involves a series of tradeoffs. As we move from performance testing to simulations to objective tests, our observations become more standardized but less realistic. As a result, our confidence in some of the assumptions involved in interpreting scores in terms of competence is strengthened, but our confidence in other assumptions is weakened. All of the approaches to evaluating professional competence have some strengths and some weaknesses.

In drawing conclusions about competence based on assessment results, at least three types of inferences are made: evaluation, generalization, extrapolation. All three of these inferences must be sound if the conclusions are to be sound. Therefore in evaluating the validity of an assessment of competence, and/or in designing an assessment procedure to generate defensible results, it is important to give special attention to the weakest links in the argument. For performance testing, evaluation and generalization are the weak links. For objective testing, extrapolation is the weak link. For simulations, any of the links may be strong or weak depending on the simulation.

It is convenient, but ultimately misguided, for advocates of performance testing or high-fidelity simulations to ignore issues of generalizability and scoring problems, just as it is convenient but misguided for advocates of

objective testing to focus their attention on the objectivity of scoring on objective tests and the generalizability of the resulting scores. We all like good news, and feel some inclination to shoot the bearer of bad tidings. But in evaluating assessment procedures, it is important to play the devil's advocate. Claims about the validity of performance test and high fidelity simulations cannot be accepted without evidence indicating that the scoring is defensible and that the results are generalizable, no matter how realistic, natural, or authentic the assessment. Similarly, claims about the validity of objective test scores as measures of professional competence cannot be accepted without evidence linking (directly or indirectly) performance on the test to performance in practice, no matter how reliable and objective the scores are. It is the weakest link that determines the plausibility of a chain of inference.

References

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

Benner, P. (1984). From Novice to Expert: Excellence and Power in Clinical Nursing Practice. Menlo Park, CA: Addison-Wesley.

Brennan, R.L. (1983). Elements of generalizability theory. Iowa City, IA: American College Testing.

Cronbach, L.J. (1971). Test validation. In R.L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, DC: American Council on Education.

Cronbach, L.J. (1980). Validity on parole: How can we go straight? New Directions for Testing and Measurement, 5, 99-108.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurement. New York: Wiley.

Elstein, A., Shulman, L., & Sprafka, S. (1978). Medical problem solving. Cambridge, MA: Harvard University Press.

Feightner, J.W. (1985). Patient management problems. In V.R. Neufeld & G. R. Norman (Eds.), Assessing clinical competence, pp. 183-200. New York: Springer Publishing Company.

Goran, M.J., Williamson, J.W., & Gonnella, J.S. (1973). The validity of patient management problems. Journal of Medical Education, 48, 171-177.

Hoffman, P. (1977). Continued competence assurance: some research and measurement considerations. In Proceedings of a national conference for evaluating competence in the health professions. New York: Professional Examination Service.

Hubbard, J.P. (1971). Measuring medical education: the tests and procedures of the National Board of Medical Examiners. Philadelphia: Lea & Febiger.

Kane, M. (1982). The validity of licensure examinations. American Psychologist, 6, 161-171.

Kane, M. (1986). The future of testing for licensure and certification examinations. In B. Plake and J. Will (Eds.), The Future of Testing. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kremer, B.K. (1991). Physician recertification and outcomes assessment. Evaluation & the Health Professions, 14(2), 187-200.

LaDuca, A., Engel, J.D., & Risley, M.E. (1978). Progress toward development of a general model for competence definition in health professions. Journal of Allied Health, 7, 149-155.

LaDuca, A., Taylor, D., & Hill, I. (1984). The design of a new physician licensure examination. Evaluation and the Health Professions, 7, 115-140.

McGaghie, W.C. (1980). The evaluation of competence: Validity issues in the Health Professions. Evaluation and the Health Professions, 3(3), 289-320.

McGaghie, W.C. (1991). Professional competence evaluation. Educational Researcher, 20(1), 3-9.

McGuire, C., Solomon, L., & Bashook, P. (1976). Construction and use of written simulations. New York: Psychological Corporation.

Melnick, D. (1990). Computer-based clinical simulation: state of the art. Evaluation and the Health Professions, 13, 104-120.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), Educational Measurement (3rd ed.). New York: American Council on Education and Macmillan.

Norman, G.R. . Defining competence: A methodological review. In V.R. Neufeld & G.R. Norman (Eds.), Assessing clinical competence (pp. 15-35). New York: Springer Publishing Company.

Page, G.G. & Fielding, D.W. (1980). Performance on PMPs and performance in practice: are they related? Journal of Medical Education, 55, 529-537.

Shimberg, B. (1981). Testing for licensure and certification. American Psychologist, 36, 1138-1146.

Stillman, P.L. & Gillers, M.A. (1986). Clinical performance evaluation in medicine and law. In R. Berk (Ed.), Performance assessment: Methods and applications (pp. 395-445). Baltimore, MD: The Johns Hopkins University Press.

Stillman, P. & Swanson, D. (1987). Ensuring the clinical competence of medical school graduates through standardized patients. Archives of Internal Medicine, 147, 1049-1052.

Swanson, D., Norcini, J., & Grosso, L. (1987). Assessment of clinical competence: written and computer-based simulations. Assessment and Evaluation in Higher Education, 12, 220-246.

Swanson, D. (1990). Issues in assessment of practice skills in medicine. Professions Education Research Quarterly, 12, 3-6.