DOCUMENT RESUME

ED 206 634 TM 810 488

AUTHOR Stiggins, Richard J.
TITLE Strategies for Optimizing and Documenting the Quality of Oral and Practical Examinations in Medical Education.
INSTITUTION Northwest Regional Education Lab., Portland, Oreg. Clearinghouse for Applied Performance Testing.
SPONS AGENCY National Inst. of Education (ED), Washington, D.C.
PUB DATE Apr 81
GRANT 400-80-0105
NOTE 14p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Allied Health Occupations Education; *Evaluation Criteria; *Medical Education; Occupational Tests; *Performance Tests; *Test Construction
IDENTIFIERS *Oral Examinations

ABSTRACT

Procedures are suggested for developing and using oral and/or practical assessment for the certification of professional competence in the health-care professions. Specific guidelines are offered for (1) deciding when to use oral or practical examinations, (2) developing sound examinations, and (3) evaluating the psychometric adequacy of developed oral and practical examinations. Evaluation criteria include appropriateness of test specifications, reliability, validity, and appropriateness of scoring procedures. Procedures are suggested for gathering, analyzing, and interpreting data relevant to each of these criteria. Responsibilities of test developers and test users are outlined with respect to each criterion. (Author)

STRATEGIES FOR OPTIMIZING AND DOCUMENTING THE QUALITY OF ORAL AND PRACTICAL EXAMINATIONS IN MEDICAL EDUCATION

Richard J. Stiggins

Clearinghouse for Applied Performance Testing
Northwest Regional Educational Laboratory
300 S.W. Sixth Avenue
Portland, Oregon 97204

April 1981

The material presented herein was developed by the Clearinghouse for Applied Performance Testing (CAPT) under grant #400-80-0105 with the National Institute of Education (NIE) of the Department of Education. The opinions expressed in this publication do not necessarily reflect the position of NIE and no official endorsement by the Institute should be inferred.

ABSTRACT

Procedures are suggested for developing and using oral and/or practical assessment for the certification of professional competence in the health-care professions. Specific guidelines are offered for (1) deciding when to use oral or practical examinations, (2) developing sound examinations, and (3) evaluating the psychometric adequacy of developed oral and practical examinations. Evaluation criteria include appropriateness of test specifications, reliability, validity, and appropriateness of scoring procedures. Procedures are suggested for gathering, analyzing, and interpreting data relevant to each of these criteria. Responsibilities of test developers and test users are outlined with respect to each criterion.

STRATEGIES FOR OPTIMIZING AND DOCUMENTING THE QUALITY OF ORAL AND PRACTICAL EXAMINATIONS IN MEDICAL EDUCATION

In instances where an examination is used to certify the professional competence of a health-care professional, the examiner has a professional and legal responsibility to do everything possible to ensure the quality of the examination. Those who are concerned with educational and professional assessment, including professional associations and testing agencies, have established and published criteria by which to judge the technical (psychometric) adequacy of an examination (APA, 1974), and decades of testing experience have allowed assessment experts to formulate guidelines of test utilization that will maximize the practical utility of examinations. The purpose of this paper is to review these guidelines and technical criteria as they apply to a specific type of examination often used to certify the competence of health-care professionals: the oral and/or practical performance examination. The review is intended to provide examiners with a checklist of factors to be taken into account in developing high-quality oral and practical tests and in verifying the technical quality of those tests. It is probable that there will be continued reliance on tests to certify professional competence in the future, and oral and practical examinations can be an integral part of that process. However, it is also probable that tests will come under increasing public and legal scrutiny. It is therefore incumbent upon examiners to gather for public review information on the quality of their tests and testing procedures.

In the oral or practical examination, the examinee is typically presented with a relatively complex set of stimulus materials intended to simulate an actual job-related circumstance within which the examineewill be expected to function once certified. The examinee's task is then to construct an original response to the simulated conditions and present that response to the examiner. That presentation might take the form of a discussion of appropriate clinical procedures, or it might require that the examinee perform some appropriate procedure or set of procedures in a manner consistent with standard professional practice. Maatsch and Gordon (1978) shed light on this type of assessment by comparing it to paper and pencil examinations:

> The use of simulations to evaluate student performance occupies the vast middle ground between highly reliable and practical multiple-choice tests and the more valid but frequently impractical individual assessment during real clinical encounters. The assessment of performance in simulated clinical encounters provides a more valid basis for evaluation of clinically relevant skills than a multiple-choice test covering the same subject matter because simulations stress the application of relevant knowledge and skills in a manner appropriate to the clinical problem or task presented. Multiple-choice examinations test the ability to recognize factual information or the ability to select the best alternative offered. The latter abilities are not called upon frequently or directly in clinical reality, so the evaluator can only assume that the student's possession of factual knowledge demonstrated on a multiple-choice test will correlate highly with the ability to apply the knowledge and other skills appropriately in clinical situations. (p. 123)

In considering the range of contexts within which oral and practical examinations (or any examination, for that matter) might be relevant, the distinction is often made between formative and summative assessment. Formative assessment is applied during instruction for the purpose of determining the status of student development so instruction can be planned to carry on that development (e.g., for a diagnosis of strengths and weaknesses or for course placement). Assessment used at the end of training to certify achievement of minimum acceptable levels of development is an example of summative student assessment. The discussion presented herein refers predominantly to the latter type of assessment--the summative assessment or certification of professional competence. It is at this point in the development of a health-care professional that the greatest care must be taken to ensure the systematic determination of skill. Though many of the guidelines provided are relevant to classroom assessment during training and should be applied there whenever possible, implementation of the guidelines will often require a time commitment, a level of assessment expertise and the expenditure of funds not available in the classroom.

In the text that follows, guidelines are offered for (1) deciding when to use oral or practical examinations, (2) developing a sound examination, and (3) evaluating the examination once it is developed. The primary emphasis in this paper is on the third point--the evaluation of oral and practical examinations--because of its potential impact on the development and use of such tests. Test evaluation criteria are described and procedures are presented for gathering, analyzing and interpreting data on the technical quality of oral and practical examinations.

Before outlining strategies for developing and evaluating the examinations, however, a note of caution must be sounded. There are several technical psychometric issues which can arise in the evaluation of oral and practical examinations which, if not carefully accounted for, can lead to erroneous conclusions regarding examinees' level of professional competence. This paper is not intended to deal comprehensively with these problems. Rather, it is intended to make the reader aware of the problems and to offer some general guidelines for their solution. Under no circumstances should the development of professional certification examinations be undertaken in the absence of sufficient assessment expertise and resources to deal comprehensively with each of the issues outlined below.

Deciding When to Use an Oral or Practical Examination

The decision to use a practical or oral performance examination in place of or to supplement a paper and pencil examination should be made only after the examiner has carefully considered a number of important assessment implications of such a decision. The most important of these is that the examiner would be opting for a relatively more subjective than objective evaluation procedure. The practical examination provides the examinee with a context within which to perform the often complex task or tasks that must be performed in professional practice. On the basis of observations of that performance, judges evaluate the level of professional competence demonstrated by the examinee. The subjectivity of the assessment arises from the application of internally held (subjective) standards to the performance by the judges.

6825A

C

Before opting for this type of subjective rating system, consideration should be given to the following differences between such a system and a more objective test (Ebel, 1978). Both oral and practical examinations typically require that the examinee construct original, relatively complex responses, while objective tests require recognition of the best response within the context of a single focused test item. Oral and practical examinations typically rely on a few exercises, each yielding complex answers, while objective tests rely on many specific answers. The quality of the scores resulting from the use of oral or practical examinations is determined by the level of training and competence of the performance evaluator, while the quality of scores obtained with objective tests is determined by the capabilities of the test item writer and the development process. In a sense, the responses to oral or practical examinations reflect the perspectives and individual characteristics of the examinee, while responses to the objective test reflect the individuality of the test item writer. Oral and practical examinations are relatively easy to construct and difficult to score, while objective tests are difficult to construct and easy to score. This factor is related to cost. The principal costs associated with objective tests are test development costs. If the test is reused, these one-time developmental costs do not recur. The principal costs associated with oral or practical examinations are scoring costs. If the test is reused, these scoring costs will always recur.

In general, oral and practical examinations are probably most useful in those situations where professional competence is best reflected in rather concise and complex sequences of behavior. They are probably most useful in contexts where such higher order mental operations as complex analysis and synthesis of knowledge and performance of skills are the object of assessment. And finally, they are only useful when the examiner has the resources available to provide qualified professionals to serve as judges of examinee performance. This point will be amplified later in the paper.

Guidelines for the Development and Administration of Oral and Practical Examinations

In order to construct technically sound oral and practical examinations and appropriate scoring procedures, there are several test development and test administration procedures which should be followed by the examiner. First, the examination exercises should be written (or prepared for oral presentation to the examinee) in such a manner as to be clearly focused and explicit. It is often useful to identify for the examinee the key points that should be included in the performance to be evaluated. However, this should not be done when such direction constitutes a cue to the examinee as to the correct response. The procedures for administering the oral or practical examination should allow sufficient time for examinees to consider, plan and present their response. In other words, the examinee must be given sufficient opportunity (direction and time) to demonstrate the required competence.

Under most circumstances, examinees should not be given a series of exercises from which to select. Examination specifications should cover those skills that are essential for professional success, and examinees should be required to demonstrate all or a systematically representative sample of those skills. A self-selected sample may not be representative and might result in the certification of an examinee whose skills are seriously deficient in crucial areas.

With regard to scoring the examination or conducting the performance evaluation, the examiner should be sure to make explicit to raters the criteria of acceptable performance, and these raters should be carefully trained to clearly indicate the examinee's status on each criterion. It is always useful to prepare and present to raters as part of the training sequence ideal responses to exercises. If possible, raters should be shown responses representing varying levels of performance, so as to make standards perfectly clear. Whenever feasible, more than one rater should be used to evaluate the performance of an examinee. Independent ratings of the same examinee's performance can then be summed or averaged, thus minimizing the chances that the particular positive or negative bias of any individual rater will lead to an erroneous conclusion regarding an examinee's level of performance. In addition, those evaluating the performance of examinees should do so ideally without any prior knowledge of the examinee's academic record or performance on other examination exercises. This reduces the probability that rater bias will be a factor in judging student performance.

For additional guidance on the development and application assessment based on simulation, the reader should refer to McGuire, et al. (1975), DeMers (1978), Broski (1978), and Maatsch and Gordon (1978).

Evaluating an Oral or Practical Examination

The Evaluation Criteria. There are basically four criteria to be considered in the evaluation of any examination, including oral and practical examinations. These are the appropriateness of the test design, reliability, validity, and appropriateness of scoring procedures.

An oral or practical examination to be used in the context of professional assessment can be considered appropriately designed if (1) the skills and knowledge to be demonstrated by examinees are clearly stated and reflect the skills that are actually a part of job performance, and (2) the examination exercises provide a realistic opportunity for the examinee to demonstrate the required skills and knowledge. An examination can be considered reliable to the extent that an examinee's scores are consistent from one administration of the test to another (over time), from one form of the test to another (across evaluators). Evidence of the validity of a test for the purpose of certifying professional competence often takes the form of a demonstration that examination performance is related to successful performance on the job. And finally, scoring procedures are considered appropriate if the procedures for establishing a pass/fail cutoff score can be made explicit and can be shown to be related to rationally derived minimum acceptable standards of test performance.

The remainder of this paper outlines procedures for addressing each of these criteria for the oral or practical examination. Sources of data on which to base the evaluation are suggested. Data collection procedures are outlined. Data analysis strategies are described. And, the nature of possible outcomes or conclusions of the evaluation process are described.

Verifying the Appropriateness of Test Design. In designing a test, there are two important considerations. First, test developers must construct a comprehensive description of the knowledge and skills to be assessed, and second, they must describe (with examples) the types of exercises to be used on the test. The verification of the appropriateness of the test design should focus on each of these components.

The developer of an oral or practical examination to be used for professional assessment can ensure the appropriateness of the knowledge and skills to be assessed by involving knowledgeable and experienced practitioners in the process of formulating the content and skill specifications. The goal is to be sure that the test will focus on or be representative of the full range of job-relevant skills and knowledge. This can be accomplished in a number of ways. The test developer can (1) conduct systematic observations of practitioner work samples and identify the relevant skills (i.e., conduct a job analysis), (2) involve experienced practitioners in an indepth discussion of the knowledge and skills that form the basis of their profession, and/or (3) generate potential lists of relevant skills and knowledge for distribution by mail to a large sample of experienced practitioners for the purpose of gathering their considered opinions regarding the appropriateness and relative importance of the skills and knowledge to be included in the specifications.

Evidence of the appropriateness of exercises can also be obtained via survey. When practitioners are asked to evaluate the knowledge and skills that are to be the focus of the test, they can also be supplied with sample exercises to evaluate in terms of (1) the extent to which they provide the examinee with an opportunity to demonstrate the required competence and (2) the extent to which the exercises represent real-world activities that are part of the actual health care environment.

Feedback from experienced practitioners regarding skill and exercise appropriateness should be taken into account in establishing final test specifications. Obviously, the desired outcome of this verification process is the conclusion that the oral or practical examination covers skills relevant to job performance. Evidence of test quality along these lines should be gathered and filed for public review if necessary.

Determining Test Reliability. A test is considered reliable if the scores it yields are consistent. In this case, consistency can have a variety of meanings. The scores may be evaluated in terms of their consistency over time (from one administration to another), across forms of the test, and/or, in the case of oral or practical examinations, across evaluators or raters of performance. Consistency over time and across test forms is often difficult to measure in oral or practical examination contexts because of the high cost and inconvenience of multiple test administrations. However, they are included in the discussion for consideration in those instances when they become feasible.

If the same test is administered twice to the same examinees with no interim instruction, the first and second scores should be about the same for any examinee. If the scores tend to differ, then factors other than examinee ability (e.g., poor exercises or unstable test administration conditions) are influencing the scores and the examiner would not know which score (if either) to rely on as an accurate estimate of examinee proficiency. In short, the test would be unreliable. To evaluate an examination along these lines, the test would have to be administered twice and the scores correlated to determine test/retest data.

An alternative means of approaching reliability is to verify the consistency of scores across forms of the test. In this case, forms of the test can be considered from two perspectives. To illustrate the first, suppose two ostensibly equivalent forms of a test were constructed to measure a given set of professional competencies. Yet the scores achieved by any given examinee on the two tests are vastly different. It might be that one set of exercises is more difficult than the other or the tests are really not covering the same material. In either case, the examiner would not know which score or form to rely on. Factors other than examinee ability would be influencing performance, rendering the test unreliable. To verify test form equivalence, both tests must be administered to the same sample of examinees and the scores must be correlated. Again, circumstances associated with the use of an oral or practical examination in the health care setting often make the collection of this type of data difficult.

However, this is not the case with the second type of test form equivalence. This second way of conceptualizing equivalent forms reliability is to focus on items or exercises that are intended to measure the same skill. In those cases where multiple exercises are included in the oral or practical examination to measure the same skill, evidence should be gathered to show that an examinee's score on such similar (parallel) exercises is approximately the equivalent. If scores across equivalent items are constant, then the elements of the test are considered internally consistent and the test is in that sense reliable. Scores on exercises within a test can be correlated to verify this internal consistency form of reliability.

Another general approach to the reliability issue, which is a crucial consideration in the oral or practical examination context, is the issue of consistency across evaluators or raters of examinee performance. If those who are to judge performance have the desired performance criteria clearly in mind (as a result of careful training) and if they are evaluating performance on the basis of those criteria, then two or more judges simultaneously observing the same examinee in a practical examination context should arrive at similar conclusions regarding examinee competence. If they do not, then some factor(s) other than examinee ability are influencing the scores. The performance evaluation procedures are rendering the scores unreliable.

Verification of interrater agreement is an indispensable part of the quality control research that should accompany the use of oral or practical examination. The simplest way to measure interrater agreement is to have two judges simultaneously and independently evaluate the same examinee's performance. If their ratings are not consistent, then factors other than examinee ability have come into play and a re-evaluation of the assessment is called for. Possible causes of inconsistency may include a lack of clear skill definition or inconsistent standards on the part of the raters. If the skills to be demonstrated have been judged to be clearly defined, then the problem is in the raters. Often a discussion of rating discrepancies by raters will reveal the reasons for disagreement, allowing for procedural revisions to eliminate differences. Such problems should be uncovered and resolved during the training of raters and prior to the actual implementation of the examination.

In sum, the test developer must verify the consistency or reliability of scores generated by any oral or practical examination being considered for use in certifying professional competence. Consistency over time and across forms can be relevant, but practical data collection problems may preclude their evaluation in health care settings. Consistency across raters--revealing the relative objectivity of the scoring procedures--is essential in all cases.

Validity of the Examination. The validity of a test is a function of both the test itself and the context within which the test is to be used. A test can only be judged valid or invalid in terms of the purpose(s) it is intended to serve. As with reliability, validity can be considered from a variety of perspectives.

One way of dealing with validity was discussed earlier. If a test is intended to measure a certain set of professional competencies and competent professionals, in fact, agree that relevant competencies are covered and that the exercises offer job-related contexts in which to demonstrate the competencies, then the intended purpose is satisfied and the test is said to be valid from a content perspective. Opinions of qualified experts regarding the appropriateness of test specifications and exercises constitute appropriate evidence of content validity.

Another more complicated and expensive way of considering the validity of an oral or practical examination in the context of certifying professional competence is to verify the relationship between performance on the examination and actual job performance. There are at least two means of conducting such a verification. One is to rely on historical data, if such data exist. If past examination performance data are available on a large group of subsequently certified and practicing professionals, then that test performance might be correlated with indicators of subsequent job success (such as supervisor ratings, professional accomplishments, years of service, job satisfaction, etc.) to reflect the extent to which test performance was predictive of job performance. A high correlation would indicate a high degree of validity for the examination process. The one limitation of this strategy is that it does not include those candidates for certification who failed the test and were not certified. Thus, due to the elimination of extremely low scorers who might have been low job performers, the true predictive power of the test will be underestimated.

A second approach to the test performance/job performance relationship is to arrange to have the oral or practical examination taken by a group of successfully practicing professionals, prior to its use with actual candidates for certification. This would provide the data needed to determine the extent to which test scores of competent professionals and passing examinees are like one another and at the same time different from those of failing examinees. In this way, the test is shown to be valid or invalid using one measure of job performance as the criterion.

For reasons of cost and simplicity, the most frequently used test validation strategy is content validation. Exploration of the relationship between test performance and job performance is often expensive, complicated and impractical. However, when appropriate historical data exist or when certified professionals agree to participate, examination of test/job performance relationship represents the most powerful test validation strategy.

Establishing the Pass/Fail Cutoff Score. Perhaps the most challenging problem facing any examiner, including those using oral or practical examinations, is the practical problem in determining the point on the score scale above which will be considered passing and below which will be considered failing. The principal reason that it is difficult to establish the cutoff score is that the score is almost always established on the basis of subjective judgment. That is, it is rarely possible to use an external job performance or other criterion to determine the test score that will predict failure. Rather, real-world circumstances always require the use of subjective judgments in establishing minimum acceptable levels of performance. Once again in this case, the subjective judgments that are most likely to be of value in setting the cutoff are the judgments of skilled and experienced professionals. Therefore, the process of selecting a minimum acceptable level of practical examination performance that is considered most defensible is one based on the pooled opinions of a broad array of skilled and experienced professionals.

Two specific procedures for accomolishing this have been outlined by Nedelsky (1954) and Angoff (1971). The simplest applications of these procedures in the context of the oral or practical examination would be to conduct a review of exercises by knowledegable experts, asking them to stipulate the level of performance on each exercise that would constitute minimum acceptable competence. Such opinions could be gathered via mail survey. Or, they might be more profitably gathered in a discussion of exercises by experts who would seek to come to an agreement on minimum level of performance. In either case, the goal is to obtain a consensus on required performance levels across experts on an exercise by exercise basis using each exercise that is to be part of the examination. The examiner can then summarize the results across exercises (such as by averaging them) to establish minimum performance levels for the test as a whole.

Though such exercise review procedures are subjective and therefore somewhat arbitrary, they are based on perceptions of those who are most familiar with professional practice. For this reason, in the author's opinion, the suggested cutoff score-setting procedures are far preferable to and more defensible than simply selecting a totally arbitrary cutoff (e.g., 75 percent correct) based on nothing more than tradition or the whim of the examiner.

Summary

In the health-care fields, oral or practical examinations based on work sample performance represent valuable tools for use in the assessment of professional competence. However, as with any type of test, the user of these types of examinations must construct sound tests and verify the technical quality of those tests.

In deciding whether to use an oral or practical examination in place of or to supplement a written test, the examiner must consider the assessment context very carefully. Where the criterion of acceptable professional practice is best represented in a series of concise and complex job-related behavioral sequences, oral or practical examinations may be of real value.

The sound development of any examination, including oral and practical examinations, requires that the test developer attend carefully to several factors. Exercises must be clearly focused and should provide the competent examinee with ample opportunity to demonstrate that competence. In addition, test administration and scoring procedures must be carefully planned and conducted to ensure fair and unbiased assessment.

Once developed, the oral or practical examination must be evaluated in terms of its content and skill specifications, reliability, validity, and pass/fail standard. The appropriateness of the test design can be verified via expert judgment of the comprehensiveness of the skills and knowledge assessed and the appropriateness of the types of exercises to be used in the test. The reliability or consistency of scores generated with oral or practical examinations can be evaluated over time, across exercises (or test forms), but is most appropriately and most oftenevaluated in terms of the degree of agreement among raters evaluating the sample of behavior. The validity of the oral or practical examination for its intended purpose can be evaluated in terms of the appropriateness of content tested and/or the relationship between test performance and subsequent job performance. And, the efficacy of the pass/fail decision can be determined through the collection of expert opinions on minimum acceptable levels of performance.

The examiner who observes these developmental and test evaluation guidelines when using an oral or practical examination and who maintains on file records verifying the technical appropriateness of their assessment procedures will be operating within the limits of acceptable professional practice. It is hoped that those who will profit most will be the examinees and the clients to whom they deliver services.

REFERENCES

American Psychological Association. Standards for Educational and Psychological Tests. Washington, D.C.: APA, 1974.

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement. Washington, D.C.: American Council on Education, 1971, pp. 514-515.

Broski, David C. Assessment of Products. In M. K. Morgan and D. M. Irby, Evaluating Clinical Competence in the Health Professions. St. Louis, Missouri: C. V. Mosby, 1978, pp. 116-122.

DeMers, Judy L. Observational assessment of performance. In M. K. Morgan and D. M. Irby, Evaluating Clinical Competence in the Health Professions. St. Louis, Missouri: C. V. Mosby, 1978, pp. 89-115.

Ebel, Robert L. Essentials of Educational Measurement (3rd ed.), Englewood Cliffs, New Jersey: Prentice Hall, Inc., 1978.

Maatsch, J. L. and Gordon, M. J. Assessment through simulation. In M. K. Morgan and D. M. Irby, Evaluating Clinical Competence in the Health Professions. St. Louis, Missouri: C. V. Mosby, 1978, pp. 123-138.

McGuire, C. H., Solomon, L. M., and Bashook, P. G. Construction and use of written simulations. New York: The Psychological Corporation, 1975.

Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, pp. 3-19.

14

6825A