DOCUMENT RESUME

ED 327 560 TM 015 978

AUTHOR Jones, Michael H.; And Others
TITLE Errors-and-Omissions Tests: A Methodology for Achieving Cost-Effective and Reliable Performance Assessments.
PUB DATE Apr 90
NOTE 39p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Boston, MA, April 17-19, 1990).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Art Teachers; Beginning Teachers; Commercial Art; *Educational Assessment; Job Performance; *Occupational Tests; *Performance Tests; Postsecondary Education; *Simulation; Situational Tests; Teacher Certification; Teacher Evaluation; *Test Construction; Test Reliability; Vocational Evaluation; *Work Sample Tests
IDENTIFIERS *Errors and Omissions Tests; Furniture Repair

ABSTRACT

In the domain of performance assessment, the errors-and-omissions (EOT) test falls between a work-sample test and a simulation test. The examinee works with a sample of material from the work environment. The correct answers and the exact criteria for acceptable performance are known in advance. For this study, EOTs were used for assessing entry level teachers in the areas of commercial art and furniture repair and upholstery. Thirty examinees were drawn from the pool of existing experienced teachers in Florida. A committee of six to eight experts in each field identified the competencies and skills needed to be an effective teacher. The committee members modified hypothetical examples of EOT tests, which were reviewed by additional experts in the fields, and used the examples in pilot testing with one teacher from the pool for the furniture repair test and three for the commercial art test. There were three raters for each test. The level of rater agreement was good. Future research will require larger samples to test the methodology proposed. Four appendices provide fictitious examples of commercial artist examinations. (SLD)

ERRORS-AND-OMISSIONS TESTS: A METHODOLOGY FOR ACHIEVING COST-EFFECTIVE AND RELIABLE PERFORMANCE ASSESSMENTS

by
Michael H. Jones, Ora Kromhout, and Robert Fleig
Center for Instructional Development
and Services, Florida State University
Tallahassee, Florida

A paper presented at the National Council
on Measurement in Education Meeting
Boston, April 19, 1990

Introduction

Occupational areas in which practitioners depend heavily on perceptual motor skills (including areas as diverse as automobile engine repair and surgery) pose special challenges for cost-effective assessment of professional competence. Test developers typically must decide whether to use cost-effective objective tests (e.g., multiple-choice tests) which often sacrifice validity in route to good reliability, or to use subjectively based performance assessments with which validity may be threatened by poor reliability.

The cost of reliable performance assessment has received special attention in recent years as teacher certification and professional licensure authorities have been inundated with legislative mandates to test in a variety of subject areas that usually have few examinees (e.g., acupuncture, precision machining, etc.). Information obtained from the National Occupational Competency Testing Institution (NOCTI) indicates that there are about 20 states now requiring certification tests for teachers in vocational subject areas. Teacher certification testing in vocational teacher testing would account for a great deal of new testing activity. For example, in the state of Florida there are approximately 122 individual vocational subject areas for which tests will be purchased from commercial vendors or for which test development will be contracted. The current proliferation of testing in professional licensure and teacher certification makes this realm especially ripe for implementation of more cost-effective assessment methodology.

Mandates for new testing programs, especially those that require performance assessments, impose a great strain on financial resources available for testing dollars. Directors of state testing programs have difficulty meeting testing mandates for subject areas where there are low numbers of examinees and the test of choice for a particular subject area (e.g. commercial art, acupuncture) is a costly performance measure. Faced with the dilemma of developing and administering tests for many subject areas with limited funds, test developers will

naturally gravitate toward using multiple-choice tests. Unfortunately, with multiple-choice tests it is difficult, if not impossible, to glean information from examinee responses about perceptual motor skills and the cognitive processes that guide movement. The traditional solution to assessing higher order thinking in a cost effective manner has been the oral interview and work sample test. Many professional certification boards[1], such as those in the fields of ophthalmology and radiology, are presently utilizing oral assessments to evaluate those clinical competencies that are gained through experience and practice. In addition, many state regulatory agencies use elaborate work sample tests to test perceptual motor skills. For example, dentists seeking to practice dentistry anywhere in the United States currently must take an examination which directly measures their ability to perform dental procedures on live patients.

In the realm of licensure and certification, the rules for determining whether to use oral interviews, objective examinations, or costly performance assessments to test professional competence are currently ill-defined. To date, standard criteria have not bee applied in establishing the methodology of assessment for particular areas, with potentially adverse impact on those who receive the services of practitioners. Unfortunately, there is no clear correlation between (1) the degree to which practitioners in a specific profession can directly affect the health and safety of the recipients of services and (2) the ~~soundness~~ directness of the assessment instrument that the profession uses to measure the competence of its own practitioners. For example, all dentists take elaborate and very expensive performance (work sample) examinations, while medical doctors and surgeons all take relatively inexpensive, objective multiple-choice examinations. One explanation for this is that for some areas (e.g., medicine), no tangible product is produced, and the actual services rendered cannot be replicated in a work

[1] The term "certification board" is used in reference to boards within professional associations, in contrast to state regulatory boards.

sample without endangering the health and safety of the patient.

The National Board of Medical Examiners has been trying to deal with the problem of conducting assessments that are more direct than traditional multiple-choice examinations. A part of the National Board Examination simulates work samples through sophisticated paper and pencil tests called clinical branching tests. These tests are designed to assess clinical diagnostic skills by having the examinee make decisions about the course of therapy for a patient. Through the use of latent imaging paper and special pens, the examinee receives feedback regarding the course (i.e., branch) of treatment he or she has chosen to pursue. The quality of the choice/branch chosen is then credited according to a predetermined scoring scheme. Additionally, the National Board of Medical Examiners is conducting research with the use of interactive videodisc technology that will allow the candidates to perform in simulated situations. Patient cases are represented on video and the candidate is required to order various tests and prescribe medications via computer simulation.

For the testing director faced with the development of large numbers of tests, the use of branching tests or interactive laser video simulations is not feasible for several reasons: (1) The measurement technology employed by such assessment approaches is very specialized, and highly technical. (2) The cost of test development for both branching and interactive videodiscs is very high, especially for videodiscs. For example, the production of simple instructional videodiscs covering relatively non-technical subject matter (e.g., high-school biology) can cost $100,000 to $250,000 per disk[2]. If branching and interactive videodiscs are eliminated as options for the testing authority, oral interviews and work sample tests remain.

The next two sections of this paper will examine some of the research that has been

[2] This estimate was provided by Dennis Thorp, Associate Director of the Center for Instructional Development and Services, who currently oversees interactive videodisc projects conducted by the Center.

done in the areas of oral interviews and work sample tests. The discussion will focus on the findings relating to the validity and reliability of these approaches.

Literature Review

Oral Interviews

Current empirical research on oral examinations for certification and licensure is limited. Most of the existing research on oral assessments has come from personnel and industrial psychology. These studies have focused on the traditional unstructured oral interview. A traditional unstructured oral interview is characterized by few, if any, scoring criteria and little standardization in terms of the questions asked of the subject. Numerous literature reviews of the unstructured oral examination/interview have been made in the last four decades (Arvey and Campion, 1982; Mayfield, 1964; Schmitt, 1976; Ulrich and Trumbo, 1965; Wagner, 1949; and Wright, 1969). This research indicates that the traditional oral interview is not a very effective assessment instrument. Meta-analytic reviews (Hunter and Hunter, 1984) show average validity coefficients of .14 for entry-level positions. Meta-analytic studies of interrater reliability were not found; however, individual studies usually show reliability coefficients below .40.

Recent research (Campion, Pursell, and Brown, 1988) designed to raise the psychometric properties of the employment interview by utilizing a highly structured set of questions and scoring criteria is encouraging. The interrater reliability coefficient reported by Campion, et al., was .76, which is good but not quite at the level of internal consistency, parallel form, and test-retest reliability coefficients associated with multiple-choice tests. Further, the process of structured interviewing presented by the authors does not address assessment of hands-on skills.

Work Sample Test

According to Siegel (1987), a work sample test or performance test involves a situation in which the person being tested performs one or more practical tasks drawn from or based on

the job itself. The research related to the work sample has been primarily in the realm of personnel psychology and industrial psychology. Virtually no research has been published on work sample testing in occupational and professional assessment, even though every state professional regulatory agency in the nation conducts or contracts for work sample tests.

Despite the enormous amount of work sample tests that are currently being used nationwide, the body of empirical research appears to be relatively small. A review of the work sample literature by Siegel (1988) reveals that the work sample generally produces validity coefficients superior to those obtained from all other forms of personnel assessment, with the exception of biographical predictors. Siegel cites a study by Asher and Sciarrino (1974) which reviews available literature through 1973 and shows that 43 percent of the validity coefficients reported were .50 or higher. Siegel (1954) developed and conducted a work sample test in drill-point grinding which used a dichotomous item-level scoring system similar to the one used for errors-and-omissions testing. The grand mean for intraexaminer consistency was 83% with a range of 64 to 100%. Siegel was apparently not satisfied with the level of agreement obtained and recommended that intraexaminer consistency be determined prior to assigning examiners to testing duty. This type of rater screening was used by the present author (Jones) to select examiners for the State of Florida Dental Licensure Clinical Examination between 1979 and 1983.

Errors-and-Omissions Tests (EOT's)

The errors-and-omissions methodology described in this paper was first used by the present author (Jones) for the dental prosthetics section of the State of Florida Dental Licensure Clinical Examination. The methodology was refined and adapted to deal with the problem of assessing the subject-area knowledge and skills of teachers seeking certification to teach in a variety of vocational areas in Florida. The testing methodology was designed to be a

compromise between an objective test and a work sample test. Specifically, it was designed to be more valid than an objective test (because the examinee was required to use "tools of the trade") and more objective than a work sample test (because the correct answers were known by test administrators in advance).

Before we discuss the development of errors-and-omissions tests for specific areas (commercial art, furniture repair and upholstery), it is important to understand where errors-and-omissions testing belongs in the performance assessment domain.

Literature on performance assessment reveals that the term "performance assessment" can address types of tests (e.g., work samples, videodisc simulations) and/or types of testing applications (e.g., personnel evaluations, writing skills assessment). To clarify the role of the errors-and-omissions test (EOT) as a performance assessment, we will consider what type of test it is as well as the applications for which it is ideally suited.

In the domain of performance assessment, the EOT falls somewhere between a work-sample test and a simulation test. It has some characteristics of a work-sample test because the examinee works with a sample of materials from the work environment. With an auto-mechanic examination, for example, an automobile carburetor may serve as the work sample. In this case, the carburetor could be presented with known defective components and the examinee's task would then be to identify the defects. It has characteristics of a simulation test because the correct answers and/or the exact criteria for an acceptable performance are known in advance.

An EOT is ideally suited for subject-area testing that involves a great deal of hands-on behavior--as with, for example, an electrician's test. Further, it is ideal for situations in which little money is available for laboratories, shops, and supplies needed to conduct the examination, and in which the test must be completed in a short period of time. Subject areas for which these situations might exist are included in the whole range of licensure and certifications tests (e.g., electrical contracting, carpentry, drafting, plumbing, commercial art, cosmetology, and

dentistry). The EOT does not require special facilities; it can be conducted in an ordinary room. All significant supplies can be brought to the room for the test administration.

Method

Subjects

For this study EOT's were used for assessing entry-level teachers in the areas of commercial art and furniture repair and upholstery. The examinees for the pilot test were drawn from the pool of existing teachers working in the state of Florida. Computer records were screened to identify those individuals who were recently certified or who were not presently certified but, because they applied for certification before a given date, would not be required to take the test. Evaluators were drawn from the pool of certified teachers who generally had 10+ years of experience. Because the population of working teachers in the vocational subject areas was very small (N < 30) the number of examinees who could be secured for validation and pilot testing was aiso small. Typically, one to three examinees participated in the pilot test.

Setting and Equipment

The room that was used for pilot testing had tables, chairs, and electrical outlets. The tests were designed so that all materials needed to administer a test could be packed into a portable container. For example, the commercial art test kit consisted of a portfolio case which contained a mechanical layout, a small drafting board, a T-square, a proportion wheel, a pica ruler, a triangle, and a small lamp. The upholstery test consisted of an upholstered footstool bisected to reveal the foundation, springs, padding, and other elements of the structure. The testing room was a typical conference room which contained chairs and standard six-foot by three-foot tables.

Test Development Procedures

Competencies and Skills Determination

The time allowed to complete the process of test development for the errors-and-omissions test was limited to two days. For each of the areas (commercial art and furniture repair and upholstery), one day was used by a committee of six to eight subject matter experts from teaching and industry to identify the competencies and skills (a form of job analysis) needed to be an effective teacher in the subject area. A competency was defined as a broad area encompassing many important skills (for example, "Demonstrate knowledge of mechanical layouts"). Within each competency area the enabling skills were listed (for example, "Demonstrate knowledge of using pica rulers"). The broad terms "competencies" and "skills" were used instead of traditional job analysis terms like "task and knowledge skills" and "abilities" because the subject-matter experts were familiar with these terms, since the state-prescribed student performance objectives were listed as competencies and skills. Indeed, the state student performance objectives were used as the foundation for generating the list of competencies and skills that were used to build the teacher subject-area tests. Pedagogical skills were not addressed in the subject-area tests because a separate test was constructed for this purpose. This process took about three hours.

Test Planning

Each committee received instructions in the use of various types of written assessment, including multiple-choice, short-answer, essay, true-false, and matching. Each type of assessment was examined to determine how it compared with the rest in terms of cost, practicality, and psychometric soundness. Various approaches to performance assessment were also discussed. After the training session, it was left to the committee to determine the type of

test best suited to assessing their subject area. An errors and omissions type of performance assessment was developed only if a committee felt it was the best to use. In certain vocational areas (e.g., machine shop), committees have chosen to use more direct forms of performance assessment. This process took about three hours.

Test Blueprint Construction

Committee members assigned weights to each of the competencies. Each weight was based on relative importance and was expressed as a percentage of the whole of all weights which totaled 100%. The group average for each competency was then computed. The percent values for each competency were then used to set the number of questions for the performance test. For example, if only 40 questions were going to be used on the performance test and competency X accounted for 50 percent of the importance, then 20 questions (i.e., errors and omissions) would address competency X. Skills within each competency were assigned by the committee a value of high or low in terms of importance. The skills that produced the highest ratings were generally the skills for which most questions (i.e., errors and omissions) were developed. This process took about three hours.

Test Development

Because of the extremely restrictive timeframe established for developing the errors and omissions tests, committee members were given hypothetical examples of test questions prepared by the Center's testing specialists and customized to fit a given subject area. These examples provided committee members (subject matter experts) with insight into how an errors-and-omissions test might look in their subject area. These hypothetical questions occasionally lacked content validity, but they fueled the committees' brainstorming process.

The examples were constructed by using a product commonly produced in a given

vocational area. For example, in the area of commercial art, a mechanical layout for color printing was used. A fictitious[3] example of an EOT using a mechanical layout has been constructed for purposes of illustration. Appendix 1 contains the hypothetical mechanical and a paper pica ruler that will be used in an audience participation exercise. A mechanical layout is composed of poster board--instead of paper, as is used in the hypothetical example--and contains the photo-ready material (e.g., print and pictures) that a commercial printer uses to make plates for printing a document. In addition to designating the final size of the printed document, the layout will have all of the required printing instructions.

A mechanical layout is a good choice for an errors-and-omissions test because, according to the job analysis, a very important component of practicing commercial art is the production and evaluation of layouts. Mechanical layouts are well suited to the application of errors-and-omissions testing because they require the use of many tools (e.g. proportion wheels, t-square, type gauges) used by commercial artists, and therefore allow for the inclusion of known errors that can be detected only by hands-on use of "tools of the trade". The key to a good errors-and-omissions test is to develop a product typically associated with the application of technical hands-on skills and include some of the most common operator-made errors.

After being presented with hypothetical examples like the one shown in Appendices 1 through 4, the subject-area committees would modify them or rewrite them to ensure content validity. The end result would be EOT tasks that would require the examinee to use common trade tools, work with a familiar product, and use various sensory modalities (e.g., sight, touch) and extensive enabling knowledge, and yet could be completed in half an hour to an hour. The process of developing the EOT took about one day.

---

[3]To ensure test security, actual examples are not provided.

Scoring and Measurement Strategy

Appendix 3 presents the instructions for determining final candidate scores for the EOT. A pass/fail decision is made by each evaluator for each error or omission detected by the examinee and a majority rule is then applied. That is, two out of the three evaluators must score an examinee's answer as passing before the examinee gets the point. It is believed that this strategy produces high intrarater agreement at the item level as well as score level across different groups of raters and across time. Because of the limited pool of examiners available for this project, only interrater score and item level agreement during one testing session was measured.

Content and Construct Validation

A committee separate from the test development committee of subject area experts from education and industry was asked to review the test which was developed. Committee members were asked to answer each of the performance test items independently and to evaluate the performance test questions (problems) in terms of their content validity and level of difficulty, and also to record any observations they wished to convey. Committee members were given the same tools or instruments that actual candidates would be given to answer questions or to complete tasks.

Test items were then critiqued by the entire committee in an open discussion. Revisions and/or decisions to discard were made by a majority decision. The development team chairperson or designee was present to respond to inquiries or to explain why certain topics or items were included in the test. This process took one day to complete for the commercial art test but was combined with a one-day pilot-testing session in the case of furniture refinishing and upholstery. Some data regarding construct validity was gathered for the commercial art test through the administration of the test to two incumbents who were new to the teaching

profession and one who had many years of experience. The underlying assumption was that if the test was a good measure of the construct of commercial art intructor, the teacher with more experience in commercial art instruction would perform better than the new teachers. Obviously, with such small sample sizes any evidence of construct validity that might be obtained would be circumstantial.

Pilot Testing

Pilot testing was conducted under simulated testing conditions. As much as possible, efforts were made to identify examinees for pilot testing who possessed qualifications similar to those of examinees who would be taking the test for certification. However, if the pool of new teachers was too small, a veteran teacher was used.

A committee member who was identified as a leader in the vocational industry was designated as test administrator. During the test, a review panel of three individuals evaluated the answers given by examinees. The only deviation the pilot test made from a real testing situation was that time did not permit the evaluators to receive training on how to evaluate the examinees' work. Under real testing conditions, evaluators are given time to familiarize themselves with the testing process. The test administrator directed the administration and scoring of the test, and also coordinated the determination of final grades.

Cost

The cost for developing EOT's is estimated at about nine thousand dollars per test using the methodology outlined in this paper. This is an estimated cost because EOT development received just a portion of a larger test development budget.

This cost would be comparable to developing a multiple-choice test of 50 to 100 items assuming a per-item cost of $50.00 to $180.00. The largest cost savings obtained through using

an EOT is seen in test administration cost. Since no special facilities are needed (as with many work sample tests), the cost is reduced and the time needed to administer the test is relatively short.

Results

Review of the data in table 1 shows that, of the thirty item level evaluations on the commercial art EOT, in 84% (i.e., 25/30) percent of the cases all three evaluators agreed on the examinee's pass-fail status regarding the item. Review of total scores given by each evaluator reveals that the greatest score deviation between any two evaluators is two points on a 10-point scale, which occurred between evaluator one and three for examinee number one. The remaining eight score-level comparisons between evaluators show that all were within one point or less, with the exception of evaluator 1 on examinee 1.

Table 1. Pass/Fail scores on a Commercial Art Errors-and-Omissions Test

	EXAMINEE								
	1			2			3		
	Rater			Rater			Rater		
	1	2	3	1	2	3	1	2	3
Examinee Answers									
1	F	F	F	P	P	P	F	F	F
2	P	P	P	P	P	P	P	F	P[1]
3	F	F	F	F	F	F	P	P	P
4	P	P	P	P	P	P	P	P	P
5	F	P	P[1]	P	P	P	P	P	P
6	P	P	P	P	P	P	F	F	F
7	P	P	P	P	P	P	F	P	P[1]
8	F	F	F	F	P	P[1]	P	P	P
9	F	P	P[1]	P	P	P	F	F	F
10	F	F	F	P	P	P	F	F	F
Total Scores	4	6	6	8	9	9	5	5	6
Examinee Final Score		6			9[2]			6	

[1]questions where discrepancies between examiners in the pass fail decision occurred
[2]represents the score of the commercial art teacher who was a veteran teacher

Table 1 shows that the veteran teacher produced the highest score on the test relative to the two less-experienced teachers. To reiterate, however, this data can only be regarded as circumstantial evidence of construct validity. The data in Table 1 also indicates that the examinees who were new to the teaching profession found the test fairly difficult since they answered only sixty percent or fewer of the items correctly.

Table 2 shows that om 85% (17/20) of the 20 item-level evaluations on the furniture refinishing and upholstery EOT, all three evaluators agreed on the examinee's pass-fail status regarding the item. Review of total scores given by each evaluator reveals that the greatest score deviation between any two evaluators is two points on a 20-point scale; this occurred

between evaluator one and the other two evaluators.

Table 2. Pass/Fail Scores by Item on Sections A and B of the Furniture Repair and Upholstery Errors-and-Omissions Test for One Examinee

	RATER		
	1	2	3
Examinee Answers			
1	P	P	P
2	F	P	P[1]
3	F	F	F
4	F	F	F
5	F	F	P[1]
6	P	P	P
7	P	P	P
8	P	P	P
9	F	F	F
10	P	P	P
11	P	P	P
12	P	P	P
13	P	P	P
14	P	P	P
15	P	P	P
16	P	P	P
17	P	P	P
18	P	P	P
19	F	F	F
20	F	P	F[1]
Total Scores	13	15	15

Final Score = 14

[1]questions where discrepancies in the pass-fail decision occurred

The level of agreement was also measured with the nonparametric Kappa coefficient. Siegel and Castellan (1988) state that the Kappa coefficient of agreement is the ratio of the proportion of times that raters agree (corrected for chance agreement) to the maximum proportion of times that the raters could agree (corrected for chance agreement). Table 3 presents the Kappa values and associated Z statistics for the data related to each examinee in

Table 1. Table 3 also presents the Kappa value and Z statistic for the data in Table 2. The alpha level was set at .01. Obviously, the small number of ratings and small examinee sample sizes would produce low statistical power for the nonparametric test used.

Table 3. Kappa Coefficients of Agreement and Associated Z Statistics

	Commercial Art			Furniture Repair and Upholstery
	Examinee			Examinee
	1	2	3	1
Kappa	.73	.70	.73	.75
Z	3.99[1]	1.63	3.99[1]	4.21[1]

[1]These values exceed the $\alpha = .01$ significance level (where $Z = 2.326$).

This data shows that the raters exhibit significant agreement. Note that the Z value for examinees 1 and 3 were higher than the Z value for examinee 2, even though the evaluators were in closer agreement regarding the answers given by examinee 2. This phenomenon is believed to be due to a degenerative form of the data. That is, with a relatively small sample size and very high agreement, the data becomes virtually invariant.[4] This in turn impacts the calculation of the estimated variance of the Kappa which is used to compute the Z statistic

$$Z = \frac{K}{\sqrt{\text{var}(K)}} .$$

[4]It was recommended by a professor of statistics to use the ocular bisector test for the data in Table 2. That is, look at it and if it hits you between the eyes then it's good! It is believed that this data fits the ocular bisector test of significance.

Discussion of the Educational Importance of EOT Assessment

The level of rater accuracy achieved in the pilot test is encouraging. Indeed, two factors imply that the present findings are conservative: first, time did not permit the evaluators to familiarize themselves with the testing process as thoroughly as would be required during a live testing condition; second, the time parameters also required the evaluators to complete evaluations at a rate faster than would be required under actual testing conditions. It is believed that near perfect rater reliability is possible under conditions where the raters are trained.

The EOT form of testing was enthusiastically selected by the commercial art and furniture repair and upholstery committees for many reasons: (1) This approach uses a dichotomous scoring system which provides reasonably objective and simple scoring criteria (i.e., the examinee's answer either corresponds to a known answer or it does not). (2) When the examinee answers the questions without aid of a list of answers to select from, measurement error due to guessing is greatly reduced. (3) Such testing eliminates the need for evaluators to make purely subjective evaluations (such as determining the relative aesthetics of a page layout, which would otherwise have been required in commercial art). In fact, it was felt that discriminating errors or omissions in a piece of work was more commensurate with the behavior of an instructor than producing a piece of work would be. That is, one does not necessarily need to have the motor skills to produce a fine piece of work in order to teach someone else to do so. (4) Dichotomous scoring at the item level circumvents the difficult process of establishing and defining anchor points along an ordinal or interval scale in order to evaluate each performance. (5) The fifth reason the committees chose to use the EOT pertains to its efficiency. For both commercial art and furniture repair and upholstery, a direct assessment of the actual production of a product would require more than eight hours to complete. The assessment process using an EOT could be completed in less than half the time.

It is important to point out that the errors-and-omissions type of assessment is simply one assessment tool of many and will not be the best choice in all circumstances. For example, if there are adequate funds for test development and administration, then a work sample would generally be preferred. However, even if a work sample is feasible, the measurement strategy used with the EOT--which involves the use of majority decisions on discrete components of the process--is recommended.

The application of the EOT methodology proposed in this paper is not limited to certification and licensure tests. It may provide a viable solution to the demand by classroom vocational teachers for assessment methodologies other than traditional objective testing modalities, such as multiple-choice tests. The EOT methodology would seem to offer hope also in the personnel assessment area, where funds and time for test administration are limited.

Future Research

One of the goals of EOT development was to establish a highly reliable assessment of performance skills without the cost and logistical problems associated with a work sample test. The collected data focused on rater agreement rather than validity studies because predictive or construct validity studies were out of the scope of the present project. However, the authors believe that rater agreement is very important and is overlooked in many professional licensure and certification tests involving performance assessments. The complex influence of test reliability and/or accuracy of raters on validity is paradoxical and always a good source for a lively debate (see Loevinger, 1954). The multidimensional nature of rater perception is not well understood and the serious advancement of knowledge in this area culminated with the research of the early psychophysicists Gustav Fectner and Ernst Weber in the 1800s and later H. Helson, L. L. Thurstone, and J. P. Guilford. The need for basic research into performance assessment strategies and the related rating scales in the area of professional licensure and teacher

certification testing is critical. If one fails to accurately assess the dentist, the root canal may abscess, or worse. If one fails to accurately assess the architect, the building or bridge may fail.

Currently, there is virtually no empirical research being reported in the area of performance assessment for professional licensure and certification. There are several possible reasons for this. One may be the litigious nature of this type of testing. That is, if a regulatory authority conducted research and published validity or reliability results that were unfavorable to a testing program, then legal exposure of the regulatory authority would increase significantly. A second reason is that measurement specialists and researchers have not been significantly involved in the development of performance assessments for professional licensure certification tests until very recently (i.e., the last 10 years). In fact, the majority of state regulatory licensure tests involving performance tests are still developed by regulatory board members who do not have any measurement training. In 1979, Kathryn Hecht of the Center for Cross Cultural Studies at the University of Alaska wrote a chapter for a book dealing with professional licensure testing. Dr. Hecht writes that while researching her topic, "Methodological Problems of Validating Professional Licensure Examinations," she "found an incredible lack of information, particularly research information." She also states that "a 1974 symposium entitled 'Validation of Professional Licensing and Certification Examinations: A Methodological, Dilemma,' was the first discussion ever sponsored by the National Council on Measurement in Education that addresses methodological problems with professional licensure exams." Dr. Hecht concluded her chapter with a warning that "if each occupation continues struggling on without serious attempts from groups such as NCME to provide integrated conceptual and methodological frameworks, solutions will remain a long way off." Amen!

References

Arucy, R.D., & Campion, J.E. (1982). The employment interview: A summary and review of recent research. Personnel Psychology, 35, 281-322.

Asher, J.J., & Sclarrino, J.A. (1974). Realistic work sample tests: A review. Personnel Psychology, 27, 519-533.

Campion, M.A., Elliott, P.D. & Brown, B.K. (1988). Structured interviewing: Raising the psychometric properties of the employment interview. Personnel Psychology, 41, 1988.

Hecht, K.A. (1979). Current status and methodological problems of validating professional licensing and certification exams. In M.A. Bunda and J.R. Sanders (Eds.), Practices and problems in competency-based education. Washington, DC: National Council on Measurement in Education.

Hunter, J.E., & Hunter, R.F. (1984). The validity and utility of alternative predictors of job performance. Psychological Bulletin, 96, 72-98.

Loevinger, J. (1954). The attenuation paradox in test theory. Psychological Bulletin, 51, 493-504.

Mayfield, E.C. (1964). The selection interview--A re-evaluation of published research. Personnel Psychology, 17, 239-260.

Seigel, A.I. (1954). Retest reliability by a movie technique of test administrators' judgements of performance in process. Journal of Applied Psychology, 38, 390-392.

Siegel, A.I. (1987). Performance tests. In Berk, R.A. (Ed.), Performance assessment: Methods and applications. (pp. 121-142). Baltimore: Johns Hopkins University Press.

Siegel, S., & Castellan, J.N. (1988). Non parametric Statistics for the Behavioral Sciences. (pp. 284-291). New York: McGraw-Hill.

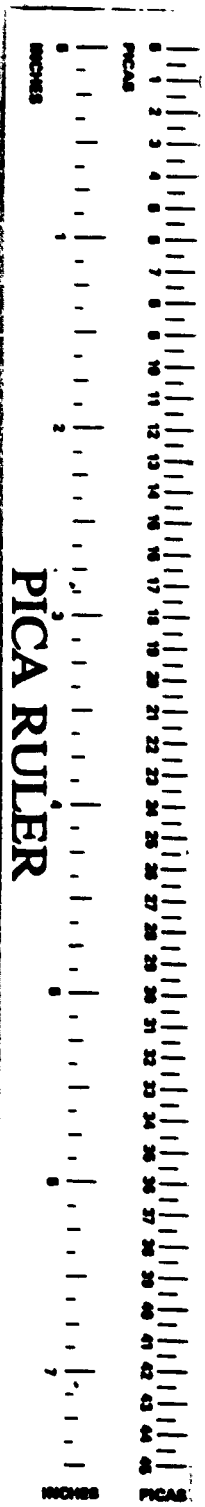
Schmitt, N. (1976). Social and situational determinants of interview decisions: Implications for the employment interview. Personnel Psychology, 29, 79-101.

Ulrich, L., & Trumbo D. (1965). The selection interview since 1949. Psychological Bulletin, 63, 100-116.

Wagner, R. (1949). The employment interview: A critical summary. Personnel Psychology, 2, 17-46.

Wright, D.R. (1969). Summary of research on the selection interview since 1964. Personnel Psychology, 22, 391-413.

Appendix 1



"O":
100% PMS 185

WE LOOK FORWARD TO HEARING FROM YOU.

OPEN ENTRIES is a participatory periodical—subscribers are also considered contributors. Take a few minutes to brag about successful materials or methodology you've developed or to express a current need. Ask for our handy, fill-in-the-blank Performance Checklist for contributions to future issues.

BLACK
½ TONE

Major funding provided by the Vocational-Technical Education Consortium of States

Produced by The Center for Instructional Development and Services, Florida State University ISSN 0739-6848

OPEN ENTRIES is promulgated at an annual cost of $12,800.00, or $0.448 per copy, to provide educators with an information system for exchange of competency-based instructional materials and methodology

"OPEN EDUCATION":
100% PMS 185
20% BLACK

OPEN EDUCATION

March 1988
Vol. 7 No. 3

The Center for Instructional Development and Services
Florida State University
Stone Building
Tallahassee, FL 32306-3019

Appendix 2

(Fictitious Example)

(CIDS) COMMERCIAL ARTIST EXAMINATION

SCORING INSTRUCTIONS

AND

ANSWER KEYS

(Fictitious Example)

General Instructions for Section A

In this section, the examinees must examine a mechanical and identify ten significant technical errors that have been made in it. Spaces are provided on the examinees' Test Booklet for Part II for the examinees to record their answers. The order in which they record their answers is not important, and the wording does not have to be exact as long as the meaning is congruent with the answers listed. At this time please familiarize yourself with the instructions found in the Test Booklet for Part II. Please pay close attention to the specifications which the mechanical was to follow.

Because the wording of the examinees' answers will vary, each evaluator will score the answers independently on an Evaluator Scoring Form to ensure maximum scoring objectivity. In order to guide you in scoring answers, an example of an "ideal answer," a "minimally acceptable answer," and an "unacceptable answer" are provided. To perform the scoring process, the evaluator must look at an examinee's answer, then scan the list of ideal, minimally acceptable, and unacceptable answers on the answer key. After you have determined whether the answer does or does not constitute a correct answer, you will simply place a check mark under the space marked Correct or Incorrect, located on the Evaluator Scoring Form.

(Fictitious Example)

Answer Key

Ideal answer. **The specifications call for the upper left paragraph to be 15 picas wide and it is 17 picas wide.**

Minimally acceptable answer. The upper left paragraph is 17 picas instead of 15.

Unacceptable answer. The 1st paragraph has wrong width.

Ideal answer. **The specifications call for the leading in the upper left paragraph to be 12 points and it is 9 points.**

Minimally acceptable answer. The leading of the upper left paragraph is 9 points instead of 12.

Unacceptable answer. The 1st paragraph has the wrong leading.

Ideal answer. **The printing instructions are written on the base art and should be on tissue.**

Minimally acceptable answer. The printing instructions should be on tissue

Unacceptable answer. The printing instructions are wrong.

etc.

etc.

etc.

.

.

.

(Fictitious Example)

Evaluator _____

Examinee I.D. No. _____

Evaluator Scoring Form

Section B

	C (Correct)	I (Incorrect)
1.	_____	_____
2.	_____	_____
3.	_____	_____
4.	_____	_____
5.	_____	_____
6.	_____	_____
7.	_____	_____
8.	_____	_____
9.	_____	_____
10.	_____	_____

Appendix 3

(Fictitious Example)

(CIDS) COMMERCIAL ARTIST EXAMINATION

TEST ADMINISTRATOR'S MANUAL

(Fictitious Example)

Determining Examinee Scores for Section A

After all of the evaluators have recorded their scores on separate Evaluator Scoring Forms, the test administrator must consolidate the scores and record the results on the examinee's test booklet in the spaces provided. In this regard, an examinee's answer will be deemed correct or incorrect based on a majority decision. For example, if three evaluators are used and two out of three mark an answer as incorrect, then the answer will be marked as incorrect on the test booklet. It does not matter how the third evaluator scores the answer.

The total score for Part I shall then be tallied and placed in the space marked Total Score for Part I on the bottom of the examinee answer form.

Appendix 4

(Fictitious Example)

(CIDS) COMMERCIAL ARTIST EXAMINATION

TEST BOOKLET FOR PART II

EXAMINEE NUMBER _____

DATE _____

DO NOT OPEN THE TEST BOOKLET
UNTIL YOU ARE TOLD TO DO SO

Copyright and Proprietary Information Protection

(Fictitious Example)

(CIDS) COMMERCIAL ARTIST EXAMINATION

Directions to Examinee

This section of the examination will assess your ability to use commercial art tools in order to identify errors in the mechanical provided. On your drawing table a T-square, a proportion wheel, a pica ruler, and a triangle have been provided.

On the photostat and acetate overlay of a mechanical provided, there are 10 significant technical errors that have been purposely made while preparing it. Your task will be to find these errors and record your answers.

When reviewing the mechanical, keep in mind the following:

- **Do not print more than one error per space provided.**
- **You must print (do not write in cursive) all answers in 20 words or less. Choose your wording carefully and be as specific as possible. For example, if the point sizes are wrong always specify the point sizes that you found.**
- **If you discover more than 10 errors, please list only the ten most important errors.**
- **You are to view the mechanical from a production perspective rather than a design perspective (i.e., style of lettering and aesthetic placement of elements).**
- **Note that there are technical specifications below which the mechanical should meet. If the mechanical does not meet a particular specification, then count it as an error. Also note that there are errors that are not related to the specifications.**

Specifications

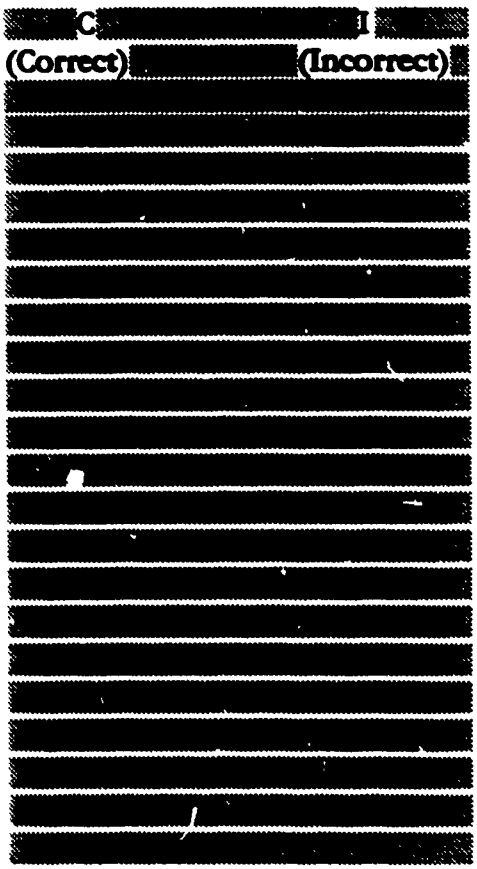
1. The paragraph in the upper left corner which begins "OPEN ENTRIES" is to be in a helvetica font and should be no more than 15 picas wide.

2. The leading for the upper left paragraph should be 12 points.

(Fictitious Example)

Answer Sheet for Section B

PLEASE DO NOT WRITE
IN THE SPACE BELOW.
THIS SPACE IS FOR
TEST ADMINISTRATOR
USE ONLY.



1. _____
2. _____
3. _____
4. _____
5. _____
6. _____
7. _____
8. _____
9. _____
10. _____