

The Reliability and Validity of a Direct Writing Assessment Program

Calvin C. Hoffman

Lisa M. Holden

Southern California Gas Company

Presented at the Annual Conference of the
International Personnel Management Association
Assessment Council, Newport Beach

June, 1997

Overview

- Writing Assessment
- Indirect vs. Direct Assessment
- Analytic vs. Holistic Scoring
- Three Case Studies:
 - Area Sales Workers (Reliability)
 - Pipeline Construction Planners (Validity)
 - Entry Management (Validity)
- Construct Validity
- Issues In Application

Writing Assessment

- Useful To Know How Well A Candidate Writes
- Can't Evaluate Via Interview
- Portfolios Or Samples Not Always Useful:
 - How Long To Prepare?
 - Who Was Author?
- Objective Measures Preferable To Testimonials

Indirect Assessment

- Multiple-Choice
- Measures Related Skills:
 - Spelling
 - Punctuation
- Advantages:
 - Good For Mass Testing
 - Cheap To Administer
- Disadvantages:
 - Does Not Tell You How Someone Writes
 - Poor Face Validity (Negative Reactions)

Direct Assessment

- Evaluate Samples Written Under Standardized Conditions
- Advantages:
 - High Face Validity
 - Samples Task of Writing
- Disadvantages:
 - Time To Administer & Score
 - Reliability?
 - Validity?

Scoring Strategies

- Holistic: Global Evaluation of Quality
- Analytic: Multiple Evaluations of Components of Written Material
- Analytic Examples:
 - Organization
 - Clarity
 - Tone
- Great Literature Review By Quellmalz (1986)
- Research Reported Here Is Based On Analytic Approach

Study One - Area Sales

- Represented Sales Job
- First Test-Based Promotional System In Company
- Highly Paid, Highly Desirable Job
- Previous Selection Based On Seniority
- Selection System:
 - Writing
 - Math
 - Oral Presentation
 - Structured Interview

Area Sales

- Writing Assessment Highly Structured:
 - Write One-Page Memo To Customer
 - 30 Minutes
 - Provided Content To Work Into Body Of Memo
- Scoring Guide Evaluated 7 Dimensions
- Behavioral Anchors At Each Scale Point (0-3)
- Extensive Rater Training

Area Sales

Statistical Analyses

- **“Classic”:**
 - Interrater (Pearson r)
 - Alpha
- **Generalizability Analyses:**
 - ANOVA Framework
 - Intraclass Correlations
 - Examine Sources of Variance

Area Sales

- Tested 259 Internal Applicants (Bidders)
- Each Memo Evaluated By Two Raters
- Classic Reliability:
 - Interrater Reliability .84 (Total Score)
 - Estimated Reliability .91 (Two Raters)
 - Alpha .73
- Generalizability:
 - Intraclass Correlation .91
 - Sources of Variance Quite Instructive

Area Sales

<u>Source of Variance</u>	<u>% Variance</u>
Raters (R)	0.2%
Dimensions (D)	19.5%
Applicants (A)	18.1%
R x D	0.2%
R x A	1.5%
D x A	39.5%
<u>R x D x A, Error</u>	<u>21.1%</u>
Total	100.1%

Planning Office (Study 2)

- Three Jobs (Union Represented)
- Multi-Method Job Analysis
- Writing Assessment Plus Four Other Predictors
- Three Criteria (Appraisal, Job Knowledge, Simulation)
- Concurrent Validation Study

Planning Office (Study 2)

- Validity Results Not Favorable For Writing Evaluation:

<u>Predictor</u>	<u>Appraisal</u>	<u>Job Know.</u>	<u>Sim.</u>
Writing	.00	.37	.17
5-Test Battery	.24	.71	.51
4-Test Battery	.33	.73	.56

Planning Office (Study 2)

- Did NOT Use Writing Assessment In Final Selection System
- Too Costly:
 - Candidate Time
 - Rater Time
 - Hassle
- Increased Costs, Decreased Validity

Ready For Management (Study 3)

- Program For Promoting Employees Into Supervision
- Company Used Tests With No Cut Scores
- Validated New Battery
- Added Cut Score
- Later Research To Add Writing Assessment

Ready For Management (Study 3)

- Concurrent Validation
- Six Criteria
- Six Predictors
- In-Basket With Writing Evaluation Were Two of Six Predictors
- Small Sample (N's From 38 to >100)
- Writing Assessment Provided Good Validity

Ready For Management (Study 3)

Average Validity Across Six Criterion Measures:

<u>Predictor</u>	<u>Validity</u>
In-Basket	.28
Writing	.36
EAS 2 (Math)	.39
PCT (Cognitive)	.29
Biodata (SPR)	.34
PET (Cognitive)	.34
Writing, SPR, PET	.43

Summary

- Demonstrated Good Reliability
- Candidates React Positively to Writing Exercises
- Validity Dependent On Job and Criterion
- Writing Assessment Seems To Tap Verbal Construct
- Balance Costs, Face Validity, Acceptance, Empirical Validity

Running head: WRITING EVALUATION

The Reliability and Validity
of a Direct Writing Assessment Program

Calvin C. Hoffman
Lisa M. Holden
Southern California Gas Company

Presented at the annual conference of the International Personnel Management Association Assessment Council, Newport Beach, June 22, 1997. For additional information, please contact the first author at Southern California Gas Company, 555 W. 5th St., Los Angeles, CA 90013.

Abstract

This series of three case studies describes a program of applied research on writing evaluation conducted in a large utility company. Two of the studies employed represented workers as subjects while the third utilized management employees. Reliability of ratings was examined several ways, including generalizability analysis, coefficient alpha, and Pearson r . Validity of writing assessments was evaluated in two concurrent validation studies using a variety of criteria, and convergence of writing assessments with a range of cognitive ability predictors was also examined. While this group of studies provides good empirical support for using direct evaluation of writing samples as a selection procedure in some settings, potential drawbacks (including cost) are also explored.

The Reliability and Validity of a Direct Writing Assessment Program

This series of studies focuses on evaluating basic writing skills in an applied selection context. While there is little personnel literature on writing assessment, published research on this topic is available in the educational literature. Efforts aimed at evaluating writing skills have often originated within the context of teaching English composition or in measuring student knowledge with essay exams. While the authors cannot exhaustively review the educational literature here, a brief introduction of issues relevant to writing assessment follows.

Indirect and Direct Writing Assessment

A major choice point in writing evaluation is that of indirect versus direct assessment of writing skills. Indirect measures typically use a paper-and-pencil, multiple-choice format. Such measures typically ask questions about spelling, punctuation and editing, but do not require assesseees to produce written text. In contrast, direct writing evaluation measures require assesseees to write original copy which is evaluated by one or more raters, typically using some type of rating scale.

Proponents of indirect formats rely on psychometric arguments to support using these measures: the tests are highly reliable, cost effective, predictive of final course grades in English, and often correlate with direct measures of writing ability. Godshalk, Swineford, and Coffman (1966) found correlations ranging from .46 to .71 between the sum of ratings of essay scores (produced by five raters) and subtests of the College Board English Composition tests. Breland, Conlon, and Rogosa (1976) found a correlation of .42 between scores on a 50-item multiple-choice test and ratings of performance on a 20-minute essay. Other researchers have found correlations falling within this range.

Critics of indirect, multiple-choice formats have argued these measures do not tap the same skills needed to perform writing production tasks (Quellmalz, 1986). Braddock, Lloyd-Jones, and Schoer (1963) regarded indirect writing measures as weak in content, construct, and ecological validity. While critics of indirect writing assessment concede that editing and comprehension abilities are adequately measured using this approach, they fault indirect measures for failing to assess such factors as unity, content, or organization (Ackerman and Smith, 1988).

In many regards, arguments about indirect versus direct measures of writing ability can be reduced to a discussion of signs and samples (Wernimont and Campbell, 1968). These authors emphasized a behavioral consistency approach where predictors should measure or sample job behavior. If reliable, valid, and cost effective direct measures of writing ability (samples) can be developed they might be preferable to indirect measures (signs) in a selection context, particularly since direct measures would likely be perceived as having greater face validity.

Direct Scoring Methods

Numerous methods for directly scoring writing skills have been developed. Quellmalz (1986) lumped criteria used to evaluate written composition into two major formats: one score or several scores. The single score may represent an overall global evaluation (often known as a holistic rating) or a rating of some feature of the composition. Formats using several scores represent ratings on a number of separate features of a composition. In some cases, one of the "features" may include a holistic evaluation of the composition along with separate ratings on individual elements.

The holistic scoring method has been used because it is relatively economical; raters can assign a global rating quickly. This approach supports the aesthetic views of some theorists that

the whole is greater than the sum of its parts (Quellmalz, 1986). Other researchers have criticized this approach since only one number is assigned and no feedback is produced to identify specific strengths and weaknesses of the composition (Quellmalz, 1986).

Methods producing multiple ratings or evaluations of a writing sample are commonly known as analytic ratings. Braddock et al. (1963) discussed a number of analytic methods including frequency counts or error counts. Counts may be made on grammatical, mechanical, or spelling errors. The present authors see over-reliance on error counts as negative since attention is focused on what is "wrong" with the composition, implying that "good" writing is simply the absence of errors. Analytic elements which are commonly evaluated include focus, organization, support, and mechanics (Quellmalz et al., 1982). Braddock et al. (1963) offered the following as possible areas of evaluation: central idea and analysis, supporting material, organization, expression (diction and sentence style), and literacy (grammar and mechanics). Both groups of researchers emphasized the need to operationally define rating elements regardless of elements used.

The literature regarding reliability of holistic and analytic scoring methods does not demonstrate a clear superiority for either approach. A study by Moss, Cole, and Khampalit (1982) included holistic and analytic scoring of two essays for students in grades 4, 7, and 10. The analytic method incorporated several elements evaluated via error counts. Interrater reliabilities for the holistic scoring method ranged from .86 to .94, while reliabilities for the analytic method ranged from .89 to .90 (all estimates were intraclass correlations and these values reflect use of multiple raters). Unfortunately, the two scoring methods did not provide a convincing level of convergent validity since intercorrelations between two methods of scoring the same essays ranged from .21 to .42 depending on sample subgroup (grade in school).

Quellmalz (1981) attempted to isolate the impact of rating method by training two groups of raters to use the same criteria. One group provided a holistic, general competence score along with diagnostic checks for component features of essays below some level of mastery. A second group provided a holistic score and separate analytic scores for each component feature. Both formats yielded agreement levels of over 90% on the general competence score. Agreement levels on component features for the combined, holistic-analytic scoring procedures (group two) were much higher than agreement on diagnostic checks given by holistic raters (group one). Quellmalz (1981) saw these results as suggesting that the requirement to assign separate scores resulted in more focused ratings.

In this series of studies we pursued a direct rather than indirect writing skills evaluation for several reasons. Along with the signs/samples viewpoint of Wernimont and Campbell (1968) discussed previously, other factors guided this decision. By requiring applicants to produce a letter, it was much easier to make a direct tie to job content, making the evaluation process more acceptable to applicants. Acceptance was a key issue since applicants in two of the three studies described here were union-represented employees bidding for jobs; lack of applicant acceptance guarantees grievances in this setting.

While the literature on reliability of holistic and analytic scoring approaches does not demonstrate a clear superiority for either method, we used an analytic approach. We believed it would be easier to obtain reliable ratings if component features of the letters were evaluated, especially if each component (or feature) was operationally defined. Additionally, a focus on component features allowed the development of "tighter," more specific anchors for scale points along rating dimensions. Finally, an agreement reached with the union during study one stipulated that anyone failing the writing assessment be given specific written feedback outlining

areas needing improvement. If a holistic rating process had been used, raters would still have had to note separate features for feedback.

The findings reported here are based on the results of three case studies. Study one describes the development and content validation of the original writing exercise and scoring protocol used in this company for a represented sales classification, and applies both classical and generalizability reliability analyses to applicant data. Study two describes the results of a concurrent validation study for represented pipeline construction planning jobs. Study three presents the results of a concurrent validation study for employees attempting to move from non-management into exempt supervisory classifications.

Method - Study One

Job and Job Analysis

The job was a union-represented sales classification (Area Sales Representative; ASR) in a large utility company. A content-oriented validation project was undertaken to develop a new selection system for screening bidders (internal, represented applicants). A multiple method job analysis was conducted to identify relevant tasks as well as knowledge, skills, abilities, and personal characteristics (KSAP's) relevant for the ASR classification.

Job analysis methods included job observation, incumbent and supervisor interviews, interviews with training staff, and development and administration of a task and KSAP inventory to incumbents and supervisors. Written communications was identified as being very important for job success (mean importance rating of 4.4 on a 5-point scale, 5 being most important). Writing was linked to job content via tasks such as correspondence with established customers, correspondence with prospective customers, and correspondence with appliance manufacturers and distributors.

Writing Exercise Development

Job analysis results guided the development of a mathematics test, oral presentation exercise, structured interview and writing exercise. The measures other than the writing exercise are not discussed further. Training staff (several of whom had previously held the job in question) provided examples of situations where sales representatives had to write original correspondence. Since the job revolved around persuading customers to use gas as an energy source, and sales representatives often provided information on appliances, these two functions were built into the writing exercise scenario.

The exercise required bidders to assume the role of a sales representative and provided a great deal of structure regarding the letter to be written. Information was provided regarding a client's name, address, and need for information regarding a specific appliance. The job analysis identified knowledge of gas appliances as highly relevant for job success, but bidders could not reasonably be expected to have knowledge of material which was trained on the job (Uniform Guidelines, 1978). The writing exercise therefore provided any appliance information which the bidder was to include in his/her letter.

Scoring Guide Development

Following the suggestions of Quellmalz (1986), the analytic scoring guide was designed to measure a number of specific elements. The scoring guide covered seven operationally defined dimensions. Specific dimensions and definitions were as follows: 1) Required Information (six sub-elements including date, bidder's name, bidder's address, client's name and address, greeting, and closing); 2) Introduction (accurately described role of job); 3) Reason for Letter (reason for writing was to introduce self and new type of appliance); 4) Request for Meeting and Follow-up (writer asked for a meeting and made provisions made to verify time and place); 5) Spelling

(extent text included misspelled words); 6) Punctuation and Grammar; and 7) Organization and Clarity (extent letter was organized, clear, and understandable). The first dimension incorporated six items scored as either present (1) or absent (0). The remaining six dimensions were each rated on a 4-point scale (0-3) with written anchors describing each scale point. The maximum possible score was 24 (six points for dimension one and three points each for the remaining six dimensions).

Rater Training and Materials

A rater training session was conducted prior to using the writing exercise. Three separate letters were created for use as training stimulus materials. An intended profile was generated for each letter which identified dimension ratings the letter was expected to capture. The three profiles were designed so (for total scores), one was acceptable (15 points), one was exceptional (22 points), and one was poor (12 points). Stimulus letters were then written to match assigned profile ratings. The process was made easier by the anchors' specificity. For example, on dimension two (Introduction), an intended rating of '1' was captured by writing the letter so no mention was made of job duties or how incumbents related to clients. That letter only included a brief mention of the job's title.

Rater Training

Four sales training staff members participated in the training session, two of whom were previous incumbents of the sales representative job. Training content briefly covered job analysis procedures and findings, and discussed how writing was applicable to the job. The writing exercise was thoroughly reviewed so raters were familiar with the problem test takers face. The writing evaluation guide was then covered in detail, focusing first on dimensions and their definitions, then discussing scale anchors and their interpretation .

After completing the review described above, rater trainees read the "average" letter and independently evaluated it using the scoring guide. Evaluation guides were collected and ratings were posted on a flip-chart along with intended ratings for the dimensions. After discussing rating discrepancies, the group repeated the process for the "exceptional" and "poor" letters. The mean correlations between intended profile and ratings provided by the four individual raters were (respectively) .93, .97, and .79. Mean correlations between all possible rater pairs (interrater reliability) for each sample letter were .91, .94, and .63 (average, exceptional, and poor letters, respectively).

Correlations for ratings on the third letter (poor) were lower than the other two letters, and requires explanation. This profile had only two dimensions with intended ratings targeted above 1.0 (elements C and D). The rating task was difficult since for most dimensions, the rater had to decide between no credit (zero rating) or slight credit (a rating of one). Given the small amount of variability over the entire profile, the smaller intercorrelations are not unexpected. In operational use, the total score assigned would be more relevant than one point reversals in ratings of separate dimensions, and total scores tracked closely to one another.

Subjects

Subjects were 259 union-represented hourly employees (173 males and 86 females) bidding for the ASR classification. All bidders completed a screening process which included the writing exercise, an applied mathematics test, and an oral presentation exercise. Bidders passing all three measures were interviewed using a structured panel interview.

Procedure

The writing exercise required bidders to read a three-page document outlining information to be incorporated into the letter they were to write. A one-page job description was provided as

extra resource material. Subjects had one hour to complete the writing exercise. Two raters independently read and evaluated each letter. The first rater marked any spelling errors with a colored pencil, so the second rater had some hints regarding how the first rater would evaluate spelling. Ratings on all other dimensions were completely independent. On all dimensions except the first, raters were allowed to make half-point ratings. Six raters in varying combinations of two provided ratings (four trained raters and the authors).

Results - Study One

Generalizability Analyses

The ratings were analyzed in a three-way crossed ANOVA design with one subject per cell. The three factors were Raters (R), Dimensions (D), and Applicants (A). Cronbach, Gleser, Nanda, & Rajaratnam (1972) refer to this design as a two facet model. Cardinet, Tourneur, & Allal (1976) refer to this design as a three facet model and discuss this apparent discrepancy in model specification. Prior to analysis, ratings for the first dimension (Required Information) were divided by two so all dimensions were on the same (three-point) scale.

Variance components were calculated using formulas provided by Cardinet et al. (1976). Table 1 provides the ANOVA summary table and variance components for all sources of variance. As suggested by Shavelson and Webb (1991) Table 1 also provides the percentage of variance attributable to each source. Raters (R) and the Raters x Dimensions (R x D) interaction each accounted for only 0.2% of variance. The largest variance component was the Dimensions x Applicants (D x A) interaction, accounting for 39.5% of variance.

Insert Table 1 about here

Generalizability coefficients were calculated using formulas provided by Cardinet et al. (1976). Table 2 provides generalizability coefficients for various combinations treating Raters

and Dimensions as either random or fixed. Raters are probably best treated as random, while dimensions should be treated as fixed. Under this condition (two raters and seven dimensions) the writing evaluation differentiated Applicants extremely well ($\rho^2=.91$). When the facet of differentiation is Dimensions, generalizability coefficients are uniformly high regardless of how Raters are treated ($\rho^2=.99$ for both models reported).

Insert Table 2 about here

Classic Reliability Analyses

Descriptive statistics and dimension intercorrelations for rater pairs are provided in Table 3. Correlations above and below the diagonal are within rater while the principle diagonal represents conspect reliability (Cattel, 1957) or interrater reliability. The correlation of raters' total scores ($r=.84$) is identical to the generalizability coefficient with one rater (random) and seven dimensions (fixed). Correcting the r of .84 for two raters using the Spearman-Brown prophecy formula provides an estimated reliability of .91, again identical to the corresponding generalizability coefficient.

Insert Table 3 about here

In operational use raters arrived at a consensus score for each dimension. Dimension consensus scores were analyzed via internal consistency and factor analysis. Treating the seven dimension consensus scores as items or scales in an internal consistency analysis produced an alpha of .73. An alpha of this moderate magnitude suggests that the writing evaluation might be factorially complex. Dimension consensus scores were factor analyzed using the SAS package (principle components with varimax rotation). A three-factor solution was found (see Table 4). The factors were named writing mechanics, content and following directions.

Insert Table 4 about here

Discussion - Study One

Generalizability and classic reliability analyses provided almost identical estimates of the writing evaluations' reliability. In either case, using two raters provided an estimated reliability for total score in the .91 range. The variance components underlying the generalizability analysis provided insights not readily apparent when using the classic reliability model.

The largest variance component was the D x A interaction, demonstrating that applicants scored differently across dimensions. Put another way, dimensions helped discriminate among applicants. The small Rater, R x A and R x D components demonstrated that raters were almost interchangeable on total score and dimension scores. The R x A and R x D interactions are potential sources of error in any measurement situation using multiple raters and multiple dimensions. A large R x D interaction would have indicated raters were rating applicants differently on the same dimensions. Raters and all associated variance components accounted for less than 2% of total variance.

Method - Study Two

Job and Job Analysis

Study two involved three jobs within the utility's pipeline planning department: Planning Assistant, Planning Aide, and Planning Technician. Data were collected as part of a concurrent, criterion-related validation study to validate a new selection system for entry into this progression. As with study one, a multiple method job analysis was completed, identifying important tasks, KSAPs, and potential criteria. Job analysis methods included reviewing written documents about the positions, interviews of incumbents and supervisors, observation of incumbents on the job, completion of Position Analysis Questionnaires (PAQ; McCormick, Jeanneret, and Mecham, 1972), and the development and analysis of a task and KSAP survey.

Survey results suggested that written communications were important, (mean importance rating of 3.8 on a 5-point scale, 5 being most important). In addition, PAQ job component validity predictions identified the Verbal construct, as measured by the General Aptitude Test Battery (GATB) as likely to be a valid predictor for these jobs (predicted validity coefficients of .19-.20 depending on classification).

Criteria

Three criterion measures were developed for this study: a research-only supervisory performance appraisal, a technical job knowledge test, and a job simulation. All KSAPs which were rated as at least moderately important during the job analysis were measured by at least one of the three criteria. The research performance appraisal consisted of 45 activities important to successful job performance. Statements were generated primarily to assess KSAPs which could not be adequately measured by the job knowledge and simulation criteria. A six-point rating scale, including “not applicable,” was provided for supervisors to indicate the extent to which each incumbent behaved in the manner described. Ratings were summed for a total score.

Again using job analysis results as a guide, subject matter experts (SMEs) generated ideas for job knowledge and simulation criteria based on typical tasks, training quizzes, and materials with which incumbents must be familiar. The job knowledge criterion asked questions in several formats, including fill-in-the-blank, calculational word problems, and short answer. The simulation criterion required incumbents to actually perform a task, such as posting revisions or calculating distance on a set of plans. Responses for both the job knowledge and simulation criteria were scored on a 0-2 scale: 0 being incorrect or blank, 1 being partially correct, and 2 being entirely correct. SMEs provided extensive assistance ensuring the technical accuracy of

both the questions and the scoring key. These criteria were piloted with several incumbents and a two hour time limit was set.

Predictors

Based on job analysis results, a test plan was developed to ensure that all KSAPs which were identified as important and needed upon entry into the progression, as well as all aptitudes found to be relevant based on the PAQ analysis, were measured. Three published tests were selected: the Adaptability test (Science Research Associates, 1942) which measures problem solving/general cognitive ability, the Industrial Reading Test (The Psychological Corporation, 1978) which assesses technical reading/verbal aptitude, and the Patterns test (Flanagan Industrial Test; Science Research Associates, 1960) which covers sketching/spatial relations and form perception.

Since published tests measuring appropriate levels of mathematics/numerical aptitude and writing/verbal aptitude could not be located, customized tests were developed by the authors. Writing exercise development is described below. A mathematics test was developed with input from SMEs regarding tasks where incumbents routinely performed mathematical calculations. Items were written so no prior knowledge of the position was required to solve the problems. The final 24-item math test could be completed within the thirty minute time limit (calculators were allowed).

Writing Exercise Development

Subject matter experts, including supervisors and staff who had previously held these positions, provided samples of letters which Planning personnel often produce, along with background information regarding what led up to the letter and what resulted. The exercise asked bidders to assume the role of a Planning Technician and write a letter to a subcontractor

requesting a change in project plans which the employee had recently reviewed. A great deal of structure and background information was provided, including the contractor's company name, contact person, and address, along with details of the circumstances necessitating the change. Although the job analysis revealed this type of request is a common occurrence for incumbent Planning Technicians, the bidders for whom the test was designed would not have access to certain facts that are trained on the job (e.g., Company policy). Information of this nature was provided in the exercise.

Along with the background facts, the exercise also listed elements to be included in the letter, such as today's date, names and addresses of the writer and the contractor, a greeting, follow-up, and closing. Employees were instructed to notify the contractor of the problem and request necessary changes in plans. They were also informed that their letter would be scored on spelling, punctuation, grammar, and organization.

Scoring Guide Development

Applying an approach similar to that used in study one, an analytic scoring guide was developed measuring six specific elements: 1) Following Directions (6 sub-elements including date, bidder's name and job title, bidder's address and phone number, contractor's name/address and contact's name, greeting, and closing); 2) Explanation of Conflict (clearly describing the problem and including all pertinent facts); 3) Request for Change (asking the contractor to alter their plans and explaining why it is necessary); 4) Follow-up (proving some specific and appropriate means to follow-up on the request); 5) Spelling (extent text included misspelled words); 6) Punctuation and Grammar (extent text included errors in grammar or punctuation); 7) Organization (extent letter was clear, organized, and understandable); and 8) Tone (extent the style of the letter was professional and considerate).

Dimensions were operationally defined so evaluators could agree on what was being rated. This exercise differed from that described in study one in that the Required elements (called Following Directions here) were previously given too much weight in the total score. As a result, the six items on the first dimension were each rated as either present (1/2) or absent (0) in this scoring guide. The remaining seven dimensions were each evaluated on a four-point (0-3) scale with written behavioral anchors describing each scale point. The maximum score was 24 (three points each for eight dimensions).

Subjects

Obtaining a sample presented some problems. There are just over 100 incumbents in the three classifications and union-represented employees are sometimes suspicious of management's intentions where testing is involved. For this reason, the second author conducted group meetings with all incumbents to explain the validation study and ask them to volunteer to take the five-test battery. The 67-person sample was representative of the entire incumbent group.

Procedure

Employees had thirty minutes to complete the writing exercise. First, they read a brief two-page document outlining information to be incorporated into the letter they were to write, then they were provided two blank sheets on which to draft their letter. Two raters independently read and evaluated each letter in a manner similar to study one, then reached consensus on each dimension. The eight dimension scores were summed into a total score on which pass or fail status would be determined.

Results - Study Two

Table 5 presents descriptive statistics and intercorrelations for criteria, predictors, and composite predictors. Predictor composites were formed by standardizing scores on each predictor and summing the resulting z -scores. Criterion intercorrelations were relatively low, ranging from .23 to .53. Predictor intercorrelations were somewhat higher, ranging from .33 to .64. Of the 15 possible validities for individual predictors, 11 were significant at the .05 level or better (one-tailed). Six of six validities for composite predictors were significant at the .05 level or better.

Insert Table 5 about here

Discussion - Study Two

Unfortunately, the writing exercise was not valid for two of the three criteria examined. While validities for the four-test and five-test composite predictors were all significant, addition of the writing exercise lowered the validity of the five-test composite predictor against each criterion measure compared to the validity of the four-test composite predictor. For this reason, the final selection battery included the other four tests but did not include the writing assessment. Keeping the writing assessment in the predictor battery would have served to decrease battery validity while significantly increasing administrative costs for candidate testing and rater scoring time. Based on the pattern of predictor intercorrelations, the writing exercise appears to measure verbal ability (r with Industrial Reading Test was .55).

Method - Study Three

Job and Job Analysis

Study three involved the utility's Readiness For Management (RFM) program, a means for promoting employees from non-management into management positions. Individuals complete a nomination process, then must pass a battery of tests before they can apply for management jobs.

Again, a multiple method approach to job analysis was used on eleven classifications representative of those passing through the RFM program. Job analysis methods included reviewing written documents about the positions, interviews of incumbents and supervisors, observation of incumbents on the job, completion and analysis of the PAQ, and the administration and analysis of a task and KSAP (Hoffman, 1987).

Criteria

Six criterion measures were available for this study: 1) company performance appraisal overall rating from 1987; 2) ratings of management skills from the 1987 appraisal; 3) overall rating on company performance appraisal from 1990; 4) salary grade in the company management job evaluation grade structure (1990); 5) change in salary grade from 1987 to 1990; and 6) a research only performance appraisal collected in 1991.

Predictors

The test battery currently used includes three measures: the Supervisory Profile Record (SPR; Richardson, Bellows, Henry & Company, 1985), a biographical inventory; the Professional Employment Test (PET; Psychological Services Inc., 1987), a measure of cognitive ability; and a writing exercise developed internally specifically for this program. All three tests were selected or developed based on a thorough job analysis (Hoffman, 1987) and validated in a pair of criterion-related validation studies (the SPR in one study; the PET and writing exercise in another; both studies are described in Holden, 1992).

Writing Exercise Development

The job analysis conducted by Hoffman (1987) recommended that an assessment of written communications be included in the test battery, so in early 1990 an exercise was developed and validated. The writing exercise actually began as an in-basket exercise. Initially,

several in-baskets developed for entry-level management were reviewed to determine the appropriate level of difficulty and typical problems. The in-basket asked candidates to assume the role of a new supervisor and deal with issues “left in the in-basket” when the previous supervisor left. Candidates were provided with instructions and background information about the hypothetical company, organization and position.

The exercise contained 14 items grouped in six problem sets. Several problems were simple (e.g., coordinate a meeting for subordinates or rearrange two meetings scheduled for the same time). One problem set involving four memos was particularly complicated. It required developing a 4/10 work schedule for 15 employees taking seniority, qualifications, and employee preferences into account. The draft in-basket was reviewed and pilot tested by several management employees, resulting in minor revisions.

Scoring Guide Development

The scoring guide was divided into three parts: the problem section (Part I) evaluated how the nominee handled each of the items, a dimension segment (Part II) evaluated the skills the nominee demonstrated in handling the problems, while Part III was similar to the writing exercise scoring guides developed for studies one and two. In the problem section, each of the ten scored items was described and a five-point scale provided. Specific behavioral anchors described the ideal response (a rating of 5), a minimally acceptable response (a rating of 3) and an incorrect or unacceptable response (a rating of 1). (Four items did not require any response). Five dimensions were assessed, including analysis/problem solving, decision making, delegation/control, interpersonal relations, and planning/scheduling/organizing. Each dimension was defined and behavioral anchors again provided for ratings 1, 3 and 5 on the five-point scales.

The writing segment of the scoring guide (Part III) was similar to those in the previous studies in that spelling, punctuation, grammar, organization, and style/tone were again assessed. These elements were again operationally defined so evaluators could agree on what was being rated. This scoring guide differed from those used previously in two ways: first, there were no “required elements” since this exercise was less structured than the two described previously, and second, the rating scale was five-point rather than four. The maximum score for the writing and dimension portions was 25 (five points for each of the five areas) and 50 for the problem section, resulting in a possible 100 points total for the in-basket exercise.

Subjects

In 1988, 80 employees participated in the SPR validation study. These individuals had previously been RFM candidates but had since been promoted into management. In addition, relatively complete appraisal histories were available in their personnel files. In 1990, these employees were asked to return to take the PET and the in-basket. Due to individuals leaving the Company and special assignments, only 60 were able to participate. In addition, 44 then-current RFM candidates volunteered to take the tests for research purposes only. Their results could be used only for predictor and adverse impact analyses, since no criterion data were available for them. In exchange, all participants received detailed feedback about their performance.

Procedure

Test sessions required approximately four hours. The PET required 40 minutes to administer, plus 15 minutes for instruction, while two and one-half hours was allowed for the in-basket. Two raters independently read and evaluated each in-basket, then discussed their ratings to come to a consensus on each item or dimension. Consensus ratings were summed within each section, as well as overall. Data from the test sessions were combined with file drawer data on

two additional predictors, the Employee Aptitude Survey mathematics test (EAS 2; Psychological Services Inc., 1952) and the Wesman Personnel Classification Test-Part I: Analogies (PCT; The Psychological Corporation, 1946). Table 6 provides means, standard deviations, predictor and criterion intercorrelations, and validities of each predictor with each criterion measure. Table 7 provides details on interrater reliability for the in-basket and writing evaluation predictors. Table 8 summarizes average validity for each predictor across the six criteria available.

Insert Tables 6, 7, and 8 about here

Results and Discussion - Study Three

Of 42 possible validity coefficients, all were significant at the .05 level or better (one-tailed). Reliability coefficients for the writing evaluation are lower than in study one since more raters were used in this study, and less rater training was provided since this was a research-only validation (study one data came from “live use” on actual candidates so it was critical that rating reliability be as high as reasonably attainable). Of the six individual predictors examined, the writing evaluation component of the in-basket had the second highest average validity across the six criteria available. A composite predictor battery of the SPR, PET, and writing evaluation provided the highest mean validity across the six criteria.

General Discussion

This study provides conditional support for the applied use of direct evaluation of writing skills and analytic scoring schemes in personnel selection. While some writers in the educational literature have made an almost Gestalt-like reference to evaluating the "whole" via global writing evaluation, study one suggests the probable existence of multiple dimensions of writing competence in what was a fairly simple writing evaluation exercise. If a global writing

evaluation had been used in this series of studies, most of the variance might have been "lost". In study one, the Dimension and D x A interaction variance components together accounted for 59% of the variance in the data set.

While raters can probably be taught to provide reliable global ratings, the lack of separate dimensions could be a handicap in an applied selection setting, especially where the writing assessment is being applied to existing employees trying to move within an organization. The analytic approach used here provided good information on which applicants were differentiated. Applicants are likely to find dimension-related feedback useful and to be able to address deficiencies noted. Narrative feedback provided to applicants was an integral part of obtaining union acceptance in this organization.

From the practitioner's viewpoint, the analytic writing evaluation method described here might be more attractive than an indirect measure in some settings. The analytic method can be developed more quickly than a multiple-choice test requiring pretesting of large samples prior to item analysis. It is also probably easier to demonstrate the content validity of a direct text production task than for an indirect measure. Finally, a text production approach has more face validity for test takers. The advantages of direct writing assessment must be balanced against the greater administrative burden which a direct writing evaluation entails, especially given the scoring time required and the time required to train raters to reach acceptable levels of reliability.

Regarding criterion-related validity, the writing evaluation did not fare as well in study two (construction planning jobs) as it did in the RFM study (study three). In study two, the writing evaluation was a significant predictor of only one of three criteria. Since the other four predictors all predicted the validation criteria better than the writing evaluation, and the writing evaluation was more administratively burdensome, it was dropped from the predictor battery

prior to implementation. From a validity standpoint, the writing evaluation performed quite well on the RFM sample (study three). Of the six predictors examined, the writing evaluation ranked second in validity, with an average validity of .36 ($p < .01$ one-tailed).

Regarding construct validity, the writing evaluation correlated most highly (.55) in study two with a measure of reading. The writing evaluation in that study correlated in the low .30 range with the rest of the predictors in that study. In study three (RFM), the writing evaluation correlated most highly (.46) with the PET, a measure of general cognitive ability. The lowest correlations between the writing evaluation and the other predictors in study three were with the PCT (.24), an analogies test which should also measure general cognitive ability, and with the SPR (.26), a biodata measure. None of these correlations are so high that they suggest the writing evaluation is measuring exactly the same constructs measured by these other predictors.

Future research should examine writing assessment in the context of different exercises. Quellmalz (1986) and Quellmalz, Cappell and Chow (1982) suggest that mode of discourse (to use the educational term) can greatly impact evaluations of writing ability. The exercise used in study one, though appropriate for the job in question, was very simple. In a more complex writing exercise, will the same scoring approach work as well? Type or difficulty of writing task could easily added as another facet in a generalizability study.

References

- Ackerman, T.A. and Smith, P.L. (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. Applied Psychological Measurement, 12, 117-128.
- Braddock, R., Lloyd-Jones, R., & Schoer, L. (1963). Research in written composition. Champaign, IL: Nation Council of Teachers of English.
- Breland, H.M., Conlon, G.C., and Rogosa, D. (1976). A preliminary study of the test of standard written English. Princeton, NJ: Educational Testing Service.
- Campbell, D.T. and Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.
- Cardinet, J., Tourneur, Y., and Allal, L. (1976). The symmetry of generalizability theory. Journal of Educational Measurement, 13, 119-135.
- Cattell, R. B. (1957). Personality and motivation structure and measurement. New York: Harcourt, Brace and World.
- Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.
- Doverspike, D., Carlisi, A.M., Barrett, G.V., and Alexander, R.A. (1983). Generalizability analysis of a point-method job evaluation instrument. Journal of Applied Psychology, 68 476-483.
- Godshalk, F.I., Swineford, F., and Coffman, W.E. (1966). The measurement of writing ability. New York: College Entrance Examination Board.
- Hoffman, C.C. (1987). Job analysis results: Readiness For Management (RFM) program. Unpublished technical report, Southern California Gas Company, Los Angeles.

Holden, L.M. (1992). Job analysis and validity study for the distribution planning office technical progression. Unpublished technical report, Southern California Gas Company, Los Angeles.

Holden, L.M. (1992). Validity of the Readiness For Management (RFM) selection system. Unpublished technical report, Southern California Gas Company, Los Angeles.

McCormick, E.J., Jeanneret, P.R., and Mecham, R.C. (1972). A study of job characteristics and job dimensions based on the Position Analysis Questionnaire (PAQ). Journal of Applied Psychology, 56, 347-368.

Moss, P.A., Cole, N.S., and Khampalit, C. (1982). A comparison of procedures to assess written language skills at grades 4, 7, and 10. Journal of Educational Measurement, 19, 37-47.

Psychological Services, Inc. (1987). Professional Employment Test. Glendale, CA: Test Publications Divisions, author.

Quellmalz, E.S. (1981). Report on Conejo Valley's fourth-grade writing assessment. Los Angeles, CA: Center for the Study of Evaluation, University of California.

Quellmalz, E.S. (1986). Writing skills assessment. In R. A. Berk (ed.), Performance assessment. Baltimore, MD: The Johns Hopkins University Press.

Quellmalz, E.S., Capell, F., and Chow, C.P. (1982). Defining writing domains: Effects of discourse and response mode. Journal of Educational Measurement, 19, 241-258.

Shavelson, R.J. and Webb, N.M. (1991). Generalizability theory: A primer. Sage: Newbury Park, CA.

Shavelson, R.J., Webb, N.M., and Rowley, G.L. (1991). Generalizability theory. American Psychologist, 44, 922-932.

Uniform guidelines on employee selection procedures (1978). Federal Register, 43, 38290-38315.

Wernimont, P.F. and Campbell, J.P. (1968). Signs, samples, and criteria. Journal of Applied Psychology, 52, 372-376.

Table 1

Analysis of Variance and Variance Components for Generalizability Analysis: Study One (Area Sales Representative)

<u>Source</u>	<u>df</u>	<u>SS</u>	<u>MS</u>	<u>δ^2</u>	<u>Percent of Variance</u>
Raters (R)	1	2.5949	2.5949	.0011	0.2
Dimensions (D)	6	426.1408	71.0235	.1349	19.5
Applicants (A)	258	649.9101	2.5610	.1252	18.1
R x D	6	3.5142	.5857	.0017	0.2
R x A	258	56.8377	.2203	.0106	1.5
D x A	1548	1071.7878	.6924	.2732	39.5
R x D x A, error	<u>1548</u>	<u>226.0573</u>	<u>.1460</u>	<u>.1460</u>	<u>21.1</u>
Total	3625	2436.8428	_____	.6927	100.1

Note: δ^2 is variance component.

Table 2

Generalizability Coefficients (ρ^2) for Writing Exercise: Study One (Area Sales Representative)

Facet of Generalization		
Random	Fixed	ρ^2
Applicants ^a		
Raters ₁ Dimensions ₇	_____	.23
Raters ₂ Dimensions ₇	_____	.26
Raters ₂ Dimensions ₇	_____	.70
Raters ₁	Dimensions ₇	.84
Raters ₂	Dimensions ₇	.91
Dimensions ₇	Raters ₂	.73
Dimensions ^a		
Raters ₂ Applicants ₂₅₉	_____	.99
Applicants ₂₅₉	Raters ₂	.99

a - Facet of differentiation

Table 3

Dimension Intercorrelations and Descriptive Statistics for Raters^a: Area Sales Representative

<u>Dimension</u>	<u>Rater 1</u>		<u>Dimension Intercorrelations</u>								<u>Rater 2</u>	
	<u>m</u>	<u>s</u>	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>	<u>m</u>	<u>s</u>
1. Required Information	2.8	.3	<u>.85</u>	.08	.07	.07	-.02	.11	.12	.24	2.8	.3
2. Introduction	1.7	.9	.04	<u>.81</u>	.46	.21	.09	.15	.25	.51	1.7	.9
3. Reason for Letter	2.4	.7	.06	.36	<u>.54</u>	.19	.19	.24	.32	.44	2.4	.7
4. Meeting and Follow-up request	2.2	.8	.05	.23	.30	<u>.78</u>	.18	.21	.32	.50	2.2	.9
5. Spelling	2.1	.9	-.02	.09	.20	.20	<u>.84</u>	.43	.34	.56	1.9	.9
6. Punctuation & Grammar	2.0	.7	.10	.23	.33	.23	.39	<u>.61</u>	.52	.57	1.9	.7
7. Organization & Clarity	2.2	.7	.04	.35	.47	.35	.42	.60	<u>.55</u>	.61	2.2	.7
8. Total	18.3	3.2	.22	.59	.66	.59	.58	.68	.78	<u>.84</u>	17.9	3.1

^a Underlined values indicate interrater reliability for raters 1 and 2; N = 259; values above and below diagonal are within rater.

Table 4

Factor Analysis Results: Study One (Area Sales Representative)

	<u>Dimension</u>	<u>Factor 1</u>	<u>Factor 2</u>	<u>Factor 3</u>
1.	Required Information	-.036	-.114	.888
2.	Introduction	.020	.863	-.003
3.	Reason for Letter	.209	.790	.090
4.	Meeting and Follow-up	.228	.257	.561
5.	Spelling	.850	-.001	-.192
6.	Punctuation and Grammar	.745	.142	.373
7.	Organization and Clarity	.690	.379	.347
	Variance Explained	38.8%	32.7%	28.6%

N = 259

Table 5
Descriptive Statistics and Intercorrelations for Criteria, Predictors, and Predictor Composite: Study Two - Planning Office

<u>Criteria</u>	<u>m</u>	<u>s</u>	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>
1. Appraisal	180.7	25.6	—						
2. Job Knowledge	82.7	15.9	.37**	—					
3. Drafting	44.3	12.0	.23	.53**	—				
<u>Predictors</u>									
4. Patterns	7.0	3.9	.32*	.59**	.47**	—			
5. Adaptability	18.9	4.9	.20	.49**	.49**	.39	—		
6. Reading	30.5	3.7	.19	.52**	.29**	.37	.47	—	
7. Math	12.3	4.6	.28*	.64**	.47**	.51	.64	.43	—
8. Writing Evaluation	16.0	3.2	.00	.37**	.17	.34	.34	.55	.33
<u>Composite Predictors</u>									
9. 5-Predictor Composite	--	--	.24*	.71**	.51**	—	—	—	—
10. 4-Predictor Composite (Less Writing)	--	--	.33*	.73**	.56**	—	—	—	—

Notes: n ranges from 60 to 67; alpha for writing evaluation is .82; alpha for math test is .83; predictor composites formed via unit weighting.

Table 6

Descriptive Statistics and Intercorrelations for Criteria, Predictors, and Predictor Composites: Study Three (Ready for Management)

Criteria	<u>m</u>	<u>s</u>	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>	<u>9</u>	<u>10</u>	<u>11</u>	<u>12</u>	<u>13</u>
1. Appraisal (1987)	4.3	0.5	—												
2. Skills (1987)	18.4	2.7	.54	—											
3. Appraisal (1990)	4.7	0.7	.32	.35	—										
4. Salary Level (1990)	10.6	1.6	.23	.30	.45	—									
5. Level Change (1987 - 1990)	1.8	1.4	.34	.30	.40	.74	—								
6. Research Appraisal (1991)	3.9	0.5	.27	.19	.55	.33	.35	—							
Predictors															
7. In-Basket	15.6	4.3	.18	.35**	.31**	.24*	.27**	.30**	—						
8. Writing Evaluation	19.0	4.1	.34**	.53**	.39**	.23*	.28**	.37**	.73	—					
9. Math (EAS2)	37.4	12.2	.20*	.34**	.43**	.40**	.50*	.44**	.41	.39	—				
10. Personnel Classification Test (Analogies)	21.8	6.4	.13*	.26*	.27*	.41**	.39	.28*	.37	.24	.47	—			
11. Supervisory Profile Record (SPR)	25.2	3.7	.34**	.41**	.29*	.46**	.38**	.16	.34	.26	.20	.30	—		
12. Professional Employment Test (PET)	25.8	7.4	.30**	.38**	.28*	.45**	.43**	.21*	.53	.46	.65	.68	.47	—	
13. Battery (Writing, SPR, PET)	?	?	.36**	.49**	.41**	.49**	.48**	.33**	—	—	—	—	—	—	—

Table 7

Interrater Reliability of In-Basket and Writing Evaluation: Study Three (Ready For Management)

<u>Predictor</u>	<u>Reliability</u>	<u>Estimated Reliability</u>
In-Basket	.79	.88
Writing Evaluation		
1. Spelling	.56	.72
2. Punctuation	.63	.69
3. Grammar	.44	.66
4. Organization & Clarity	.34	.51
5. Style/Tone	.48	.65
Writing Total	.61	.76

Note - estimated reliability based on Spearman-Brown prophecy formula for two raters.

Table 8

Average Validity of RFM Predictors Across Six Criteria: Study Three

<u>Predictor</u>	<u>Average Validity</u>
In-Basket	.28*
Writing Evaluation	.36**
Employee Aptitude Survey 2 (Math)	.39**
Personnel Classification Test (Verbal)	.29**
Supervisory Profile Record (SPR)	.34**
Professional Employment Test (PET)	.34**
Battery (Writing Evaluation, SPR, PET)	.43**