Center for
Effective
Organizations

A COMPARISON OF CRITERIA FOR TEST VALIDATION: A META-ANALYTIC INVESTIGATION

CEO PUBLICATION
T 87-17 (108)

BARRY R. NATHAN
University of Southern California

RALPH A. ALEXANDER
University of Akron

MARCH 1994

Center for Effective Organizations - Marshall School of Business
University of Southern California - Los Angeles, CA 90089–0871
(213) 740-9814 FAX (213) 740-4354
http://www.marshall.usc.edu/ceo

A Comparison of Criteria for Test Validation: A Meta-Analytic Investigation

CEO Publication
T87-17 (108)

Barry R. Nathan
University of Southern California

Ralph A. Alexander
University of Akron

Center for Effective Organizations
School of Business Administration
University of Southern California
Los Angeles, Ca 90089-1421
(213) 743-8765

A COMPARISON OF CRITERIA FOR TEST VALIDATION: A META-ANALYTIC INVESTIGATION

Abstract

Meta-analyses of validity coefficients from tests of clerical abilities for five criteria; supervisor ratings, supervisor rankings, work samples, production quantity, and production quality, were conducted and the resulting expected true validities were compared. Ratings, rankings, work samples, and production quantity all resulted in high test validities. Validities resulting from ratings and quantity of production criteria were highly similar across tests. Validities resulting from rankings and work samples were on the average higher than those from ratings and quantity of production. The fifth criterion, quality of production, had low predictability and did not generalize across situations.

Author Notes

This paper is based on the first author's doctoral dissertation, directed by the second author. An earlier version of this paper was presented at the 28th annual conference of the Midwest Academy of Management, April, 1985, in Urbana-Champaign, IL. We would like to thank Dr. Kenneth Pearlman who most graciously shared his data and his time with the first author.

One of the most important decisions made when conducting a validation study involves the choice and development of an appropriate criterion. While technical and legal standards require that the measures of work behavior or performance used as criteria for validating a test be determined from a carefully conducted job analysis (Nathan & Cascio, 1986), personnel specialists may still have to choose among different measures of work performance, e.g., supervisory ratings, production output, work samples, all of which may be job-related. If different types of job-related criteria exist, then selecting criteria on the basis of expediency or availability would seem justifiable. On the other hand, Brogden and Taylor (1950) argued that criteria should be chosen that facilitate the validation process, i.e. result in the highest validity coefficients. In addition to these practical concerns, investigating differences between types of criteria could also lead to a better scientific understanding of our performance measures (Schmitt & Schneider, 1983). This question of whether different criteria result in different validity results for the same tests, i.e. are differentially predictable, was the focus of this investigation.

Ideally, all criteria judged to be job-related should correlate with valid tests; in practice, criteria can differ in their predictability due to differences in contamination and deficiency. Production indices can be deficient to the degree that important aspects of the job are not measured. For example, typing productivity would be a deficient measure of a secretary's overall job performance. Likewise, the same production measure could be contaminated by factors beyond the secretary's control, such as equipment, e.g. typewriter versus a word processor, or the supervisor's assignments, e.g. typing letters versus

statistical reports. Supervisor's ratings too, can be deficient or contaminated to the degree certain dimensions of the job or even specific critical incidents of behavior are under- or over-emphasized relative to the employee's total performance. In addition, ratings also may be deficient or contaminated by such rating errors in judgment, as halo, leniency, contrast effects, etc., deficiencies in observations, and of particular concern over the past twenty years, intended or unintended bias due to age, race and sex. As a result, many possibilities exist that could cause differential predictability between ratings and objective production data.

Surprisingly, very few validation studies report result from different criteria. In fact Lent, Aurbach, and Levin's (1971) review of 467 studies in the Validity Information Exchange of Personnel Psychology from 1954-1965 found only eight studies in which more than three criteria were used, and an average number of criteria per study of only 1.15. Recently, Schmitt and Schneider (1983) reviewed 98 studies reported in the Journal of Applied Psychology and Personnel Psychology from 1965 to 1978, and found validities from objective and subjective criteria to be virtually identical, averaging .29 and .31, respectively.

Following Schmitt and Schneider's (1983) recommendation that more work in this area be done, criteria predictabilities were investigated by conducting meta-analyses of different criteria used to validate the tests of clerical worker performance collected by Pearlman (Pearlman, 1979; Pearlman, Schmidt, & Hunter, 1980). The five criteria compared in these analyses included two soft criteria, supervisory ratings and supervisory rankings; two hard criteria, output quantity and quality of output; and work sample performance.

METHOD

Data

Data used in this study consisted of a subset of 2795 validity coefficients for proficiency criteria collected by Pearlman (Pearlman, 1979; Pearlman et al., 1980). The Pearlman data set was based on an extensive search of published and unpublished validity studies pertaining to clerical jobs. Although ten test types and eight criterion types were recorded in this data set, many of the test-criterion combinations included very few correlation coefficients. As a result, four criteria were eliminated from these analyses: hired/not hired, wages, job level, and any composite of the various proficiency measures. Also three test types were eliminated: reasoning ability, performance tests, and clerical aptitude tests. A breakdown of the remaining 1709 validity coefficients into test-criterion combination is presented in Table 1. Of the 511 coefficients where sex was clearly identified, 90% were from female samples. The data represent validity coefficients combined across all clerical jobs and job families.

Insert Table 1 About Here

For the studies using hard criteria, production measures were reviewed and recorded as either production quantity (total output), or production quality (errorless production, error-corrected production, or amount of errors).[1] Rankings consisted of any subjective comparative evaluation and included simple rankings, rank-order ratings, paired comparison ratings, and keep/retain rankings. Validity generalization results reported by Pearlman et al. (1980) and the job family analysis conducted by Schmidt, Hunter, and Pearlman (1981) indicated that the

validities across all clerical jobs are similar, and further separation is unnecessary.

Data Analysis

The validity generalization procedures outlined by Schmidt et al. (1980) was used to estimate the true predictability of each criterion (as opposed to the validity of a test). For each test-criterion combination, validity coefficients were combined, and the resulting distributions of rs were corrected for test unreliability, sample size, and range restriction. No corrections were made for criterion unreliability since such unreliability contributes to the predictability of the criterion, the concern of this study. True criterion predictability was estimated by the mean of the corrected prior distribution. Ninety percent credibility values of these distributions were also calculated. In order to estimate the variance in studies due to test unreliability and range restriction, Schmidt and Hunter (1977) and Pearlman et al. (1980) relied on reasonable assumed distributions of these effects. The same distributions were used for estimating these effects in this study. If type of criterion makes a difference in the calculation of test validity, we would expect large differences in average validities and 90% credibility values.

Only five studies were found in which more than one criteria were collected on the same sample. It was felt that a meta-analysis on such a small number of studies would give unreliable results and very tenuous conclusions. Thus the meta-analytic Z-test (Ford, Kraiger, & Schectman, 1986; Rosenthal & Rubin, 1982) that tests whether there is a statistically significant difference between effects across studies was not conducted. Instead, estimates of the percentage of overlap (Tilton,

1937) between distributions of validity coefficients from different criteria were calculated. In the absence of an acceptable significance test, this analysis provides an additional indication of the degree of similarity or difference that could be expected in choosing one criteria over another when conducting a validation study.

RESULTS

An overall meta-analysis for each test type, ignoring criterion differences, indicated that statistical artifacts alone (sampling variability, unreliability, and range restriction) did not account for all of the variability in validities across studies. Chi-square tests for significant variance beyond that due to chance (Hunter, Schmidt, & Jackson, 1982) yield the following significant results for each test type across all criteria: general mental χ^2 (191, n=14,977) = 1,293.66; p<.001; verbal ability, χ^2(350, n=28,711) = 1526.94, p<.001; quantitative ability, χ^2(358, n=28696) = 755.70, p<.001; perceptual speed, χ^2(358, n-37,507 = 1,100.28, p<.001; memory, χ^2(99, n-6,734) = 253.81, p<.001; spatial/mechanical ability, χ^2(99, n=8,010) = 245.12, p<.001; and motor ability χ^2(128, n=10,580) = 418.56, p<.001. Similarly, variance due to statistical artifacts alone (41%, 48%, 69%, 51%, 61%, 58%, and 56%, respectively, for each test type listed above) was sufficiently small to justify investigating the effect of each criterion separately.

Criterion Predictability

The question of whether all criteria are predictable for all tests was investigated by inspecting the means and 90% credibility values of the prior distribution of predictability coefficients. In a few cases

statistical artifacts accounted for more than 100% of the total variability, indicating statistical artifacts could be responsible for all of the observed variability from study to study. It is possible, of course, that for any one study we could be overcorrecting (or for that matter undercorrecting) for the effects of statistical artifacts. To be conservative, i.e. to allow for the possibility that we are overcorrecting, we assumed that at a maximum, 90% of the observed variance would be attributable to statistical artifacts (Nathan & Alexander, 1985). The standard deviations and 90% credibility values of the distribution were recalculated accordingly. Table 2 shows that for all tests, the best estimate of criterion predictability, the mean of the corrected distribution, was at least reasonably high for all but one criterion. The lone exception was production quality which had the lowest predictabilities for each test, with _rs ranging from -.01 to .32. The work sample and ranking criteria were the most predictable criteria with average estimated validities ranging from .61 to .43 and .66 to .23 across tests, respectively. The ranges of average predictabilities of ratings and production quantity were also quite similar, but more moderate in size with average estimated validities across tests of .44 to .24 and .44 to .30, respectively.

Insert Table 2 About Here

At the 90% credibility value, the key concern is whether the value is greater than zero. If it is, then we can be 90% confident that even when a conservative estimate of predicted validity is considered, the validity of that type of test using that particular criterion, generalizes across situations. Table 2 shows that for ratings,

rankings, work samples, and production quantity, every test/criterion combination but one, motor ability tests with ranking criteria, has a 90% credibility value that is greater than zero. Thus with that one exception, we can be at least 90% confident that these tests will be predictive of job performance across situations, regardless of the choice of these four criteria. Likewise we can have no confidence that using production quality as a criterion will yield valid test results since six 90% credibility results are negative and the seventh is only .01.

To summarize the results thus far, the best estimate of a tests true validity, the mean of the corrected distribution of validity coefficients, is high for four of the criteria: ratings, rankings, work samples and production quantity. The 90% credibility values vary, but more importantly, with the exception of the motor ability/ranking results, are all above zero and typically above .10. Only production quality consistently yields validity results that are low and do not generalize across tests or situations.

Overlap Analyses

Estimates of the percentage of overlap between validity distributions from different criteria are presented in Table 3. Dunnette (1966) has suggested two somewhat arbitrary "rule of thumb" cutting points for interpreting overlap, 75% and 50%.[2] Overlap greater than 75% indicates little useful differentiation; values below 50% represent unusually high differentiation between criteria.[3] Because the corrected distributions (above the diagonal) have artifactual variance due to sampling error, unreliability, and range restriction removed, the spread of these distributions and thus their overlap is reduced relative

to the observed validity distributions (below the diagonal). It therefore provides a more conservative test of whether the validity distributions from different criteria are identical.

Insert Table 3 About Here

For the observed validity distributions, the majority of the comparisons among the four highly predictable criteria: ratings, rankings, work samples, and production quantity are above 75% indicating little useful differentiation among criteria, and none are less than 50%. The overlaps between validity distribution from the two most common criteria, subjective ratings and objective production quantity, are consistently large, ranging from 80% to 97%. Even after removing the effects of statistical artifacts, the more conservative corrected distribution overlaps between ratings and production quantity are greater than 75% for five of the seven test types, and only slightly lower for the remaining two, 73% and 69%, for memory and motor ability tests, respectively. However, in general, greater variability exists in the overlap among these four criteria when the effect of statistical artifacts are removed from the validity distributions.

In contrast, the overlap of observed validity distributions for the quality of production criterion with other criteria vary a great deal depending on test type and criteria. However, once availability due to statistical artifacts are removed, a very clear pattern emerges: only five of the 28 comparisons between production quality and the other four criteria are large, and 18 (64%) are below the 50% cutpoint indicating unusually high differentiation.

In summary, ratings and quantity of production have overlapping validity distributions of such a magnitude that little meaningful differentiation between the use of these criteria could be determined. This is true even after the variability due to sampling error and attenuation had been removed. Thus there is no apparent reason why personal researchers should expect "better", i.e. higher validity to result from "objective" quantity output measures of performance. This is not true in the case of quality of production criteria where the results were quite erratic across test types, and when statistical artifact variance was removed, highly differentiated from the other criteria. Furthermore, as shown in Table 2, the dissimilarity of production quality is due to its substantially lower predictability than that of the other criteria. Thus, from the practical perspective of selecting a criterion for test validation, the use of a quality of production criterion should, until further research is conducted, be avoided.

The remaining two criteria, rankings and work samples, in general showed reasonably large overlap between themselves, and because of the higher validity that often occurred with their use, more moderate overlap with ratings and production quantity. However, since these two criteria have the same or higher predictability than ratings and production quantity, the use of either of these criteria in place of ratings or production quantity would generally result in similar or higher validity results.

DISCUSSION

In the present study, the use of meta-analysis, and in particular, validity generalization, addresses some of the concern researchers have

had in regard to the choice of criteria for test validation. Four of the five criteria investigated: supervisor ratings, supervisor rankings, work samples, and quantity of production, all resulted in large expected true validity coefficients for all tests. Two of these criteria, ratings and quantity of production, produced validities that were consistently highly similar. Thus, there is no support for the assumption that "objective" measures of performance are more predictable, than subjective evaluation (Toops, 1944). This was true for highly intellectual measures (general mental ability, quantitative and verbal abilities, and spatial/mechanical ability), more general cognitive processing measures (memory and perceptual speed), and even motor ability measures.

In fact, the "subjective" versus "objective" distinction may be more illusory than real. As Smith (1976) has noted, production measured as a function of a "standard of performance" will be subjective to the degree that choice of the standard is a subjective decision. Conversely, "subjective" evaluations are made (with some error) on the basis of observed behaviors and actual performance (Nathan & Lord, 1983). As a result, it is not unreasonable to find that both kinds of criteria can be similarly predictable.

The two other predictable criteria, rankings and work samples, tended to result in higher average validities than those of quantity of production and ratings, and therefore had less overlap in their distributions. In retrospect, this should have been expected. For example, while both rankings and ratings rely on subjective evaluations of performance, they differ in the shapes of their distributions of scores; rankings result in less restriction of range and are therefore

less likely to result in artifactually attenuated predictability coefficients. Work samples would likely be less susceptible, though not immune, to rating errors than are supervisor ratings, and are less affected by various situational or equipment factors than are production records.

Our results should not be interpretted as suggesting that objective and subjective criteria are similarly susceptible to bias. A meta-analysis by Ford et al. (1986) found the average difference between white's and black's performance across studies to be greater when performance was measured by subjective measures than objective indices, though whites performed higher on the average with either performance measure. A meta-analysis by Waldman and Avolio (1986) found the average correlation across studies between age and performance was negative for rating and positive for productivity. Certainly the opportunity for bias in objective indices of performance may be less than in subjective ratings (though not totally eliminated). On the other hand, the difference between studies could be due to the use of different occupations. Ford et al.'s samples included firefighter, police officers, bank tellers, skilled technicians, production workers, and clerical workers. Waldman and Avolio's samples included scientists, air traffic controllers, management faculty, research chemist, garment workers, and others. Ours were restricted to only clerical occupations.

Possibly the most interesting finding of this study was the consistently low predictabilities of production quality. Given the recent emphasis on quality voiced by many business and government leaders, this finding is particularly disconcerting. One explanation is that the low predictability reflects an inability to adequately or

consistently operate the construct of quality. For example, in some of our studies quality was operationalized as the total number of errors; in others, quality was computed as a function of total production output, such as total minus errors or total minus a weighting of errors. Some studies defined quality in terms of errors per unit of time.

Quality also may be a less predictable because of low reliability. The length of time used to collect production data, and the interval of time between these periods can have a considerable effect on the reliability of any production data (Rambo, Chomiak, & Price, 1983; Rothe, 1978). The problem is considerably greater when errors are considered. Since errors are likely to be relatively rare events within the context of total output, extremely long periods of time may be necessary before reliable individual differences in error-prone performance can be discerned. If this is not taken into consideration, then error-related criterion measures may be highly inaccurate and result in diminished or even negligible predictability, as was found in this study.

Another problem with the quality criterion is whether or not quality is under the employee's control at the time the measurement was taken. Deming (1975), for example, estimates that 85% of the faults in production are system related (i.e., management controlled). Only 15% can be attributed to the worker or to a machine. Even if these estimates are conservative, there is still very little variance in quality of production which would be attributed directly to employees, and therefore very little which could have been predicted by individual difference measures. It may be that American supervisors are in error

when they attribute much of our quality problems to workmanship and workforce (Garvin, 1986), certainly our results do not support this perception. Instead, the perspective of Japanese supervisors, that poor quality is the result of product design and purchase parts and materials (Garvin, 1986) should be considered. Clearly, further research into the meaning, measurement, and cause of quality is needed if it is to be reliably predicted.

While most of the conclusions reached in this investigation are quite clear cut, e.g., the similarity in predictabilities of ratings and quantity of production data, and the dissimilarity and generally poor predictability of quality of production, a number of limitations must be kept in mind. First, as noted previously, the data consists solely of clerical occupations in which most performance outcomes are easily operationalized and observed. Whether the same results would occur in more complex jobs such as management positions where responsibilities are more abstract and determing accountability is more complicated, awaits further investigation. Second, the number of validity coefficients available for this study was quite limited for some of the criteria and in some cases, the addition of one or two very large or very small coefficients could have had an appreciable effect on the prior distributions. Furthermore, studies using objective and subjective criteria from the same sample were not available in sufficient number to allow direct comparisons between criteria, clearly more research of this type would be desirable. However, in light of the potential sampling variability which could have occurred as a result, the consistency of our findings across different test types is impressive. Large, systematic contamination and deficiency associated

uniquely with each criterion was not, with the exception of production quality, contributing to differential criterion predictability. Had systematic criterion contamination and deficiency been present, predictabilities would have been quite different for different criteria and overlap would have been severely reduced. The present investigation found just the opposite, consistently large or moderate-to-large overlap among four of the criteria.

Finally, it must be emphasized that the study undertaken here is limited to the question of criterion predictability in a restricted sense--as dependent measures in criterion-related validation studies. This is the question of "validity extension" raised by Mosier (1951) some 30 years ago. A very different set of questions, not addressed by this study, has to do with whether different criteria are equally appropriate for other purposes, e.g., understanding job performance, feedback and employee development, merit pay and promotion decisions, etc. In general however, the results found in this study indicated that choosing a criterion for validation research may not be as serious a problem as has been generally assumed.

Footnotes

[1]Other differences in the coding of studies for this study as compared to that by Pearlman et al. (1980) can be obtained from the first author. These differences in data coding had only the most trivial of effects on the estimates of expected true validity distributions for types of tests reported by Pearlman et al. (1980), and absolutely no effect on their conclusions.

[2]Dunnette suggested Tilton's overlapping statistic as another measure of the validity of a test. The higher the overlap between test score and job performance score distributions, the higher the validity. It therefore had a different purpose than that used here, i.e., calculating overlap between validity distributions. However, since the cut-off values suggested by Dunnette are not unreasonable they were retained for purposes here.

[3]Curtis and Alf (1968) point out that where the distributions are based on unequal sample sizes, as in the case where supervisory ratings are compared to other criteria, the calculated percentage of overlap is less than would have occurred had the distributions been based on equal sample sizes. Thus our results should be considered as conservative estimates of the degree to which validity distributions from different criteria give similar results.

REFERENCES

Brogden, H. E. & Taylor, E. K. 1950. The theory and classification of criterion bias. Educational and Psychological Measurement, 10: 159-187.

Curtis, E. W. & Alf, E. F. Sr. 1968. A correlational approach for mesuring overlap of distributions and proportions of miscalculation. Psychological Bulletin, 70: 626-630.

Deming, W. E. 1975. On some statistical aids toward economic production. Interface 5(4): 1-15.

Dunnette, M. D. 1966. Personnel selection and placement. Belmont, CA: Brooks/Cole.

Ford, J. K., Kraiger, K., & Schechtman, S. L. 1986. Study of race effects in objective indices and subjective evaluations of performance: A meta-analysis of performance criteria. Psychological Bulletin, 99: 330-337.

Garvin, D. A. 1986. Quality problems, policies, and attitudes in the United States and Japan: An exploratory study. Academy of Management Journal 29: 653-673.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. 1982. Meta-analysis: Cumulating research findings across studies. Beverly Hills, CA: Sage.

Lent, R. H., Aurbach, H. A., & Levin, L. S. 1971. Research design and validity assessment. Personnel Psychology, 24: 247-274.

Mosier, C. I. 1951. Symposium: The need and means of cross-validation. I. Problems and designs of cross-validation. Educational and Psychological Measurement, 11: 5-11.

Nathan, B. R., & Alexander, R. A. 1985. An application of meta-analysis to theory building and construct validation. An example using the Miner Sentence Completion Scale. Proceedings of the 45th Annual Meeting of the Academy of Management, San Diego, CA.

Nathan, B. R., & Cascio, W. F. 1986. Technical and legal standards for performance assessment. In Berk, R. A. (Ed.), Performance assessment: Methods and Applications, Baltimore: John Hopkins Press.

Nathan, B. R., & Lord, R. G. 1983. Cognitive categorization and dimensional schemata: A process approach to the study of halo in performance ratings. Journal of Applied Psychology, 68: 102-114.

Pearlman, K. 1979. The validity of tests used to select clerical personnel: A comprehensive summary and evaluation (Technical Study TS-79-1). Washington, D.C.: U.S. Office of Personnel Management, Personnel Research and Development Center (NTIS No. PB 80-102650).

Pearlman, K., Schmidt, F. L., & Hunter, J. E. 1980. Validity generalization results for tests used to product job proficiency and training success in clerical occupations. Journal of Applied Psychology, 65: 373-406.

Rambo, W. W., Chomiak, A. M., & Price, J. M. 1983. Consistency of performance under stable conditions of work. Journal of Applied Psychology, 68: 78-87.

Rothe, H. E. 1978. Objective rates among industrial employees. Journal of Applied Psychology, 63: 40-46.

Schmidt, F. L., Gast-Rosenberg, I., & Hunter, J. E. 1980. Validity generalization results for computer programmers. Journal of Applied Psychology, 65: 643-661.

Schmidt, F. L., & Hunter, J. E. 1977. Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62: 529-540.

Schmidt, F. L., Hunter, J. E., & Pearlman, K. 1981. Task differences as moderators of aptitude test validity in selection: A red herring. Journal of Applied Psychology, 66: 166-185.

Schmitt, N., & Schneider, B. 1983. Current issues in personnel selection. In Rowland, K. M., and Ferris, G. R. (Eds.), Research in Personnel and Human Resources Management, Vol. 1, Greenwich, CT: JAI Press.

Severin, D. 1952. The predictability of various kinds of criteria. Personnel Psychology, 5: 93-104.

Tilton, J. W. 1937. The measurement of overlapping. Journal of Educational Psychology, 28: 656-662.

Toops, N. A. 1944. The criterion. Educational and Psychological Measurement, 4: 271-297.

Waldman, D. A. & Avolio, B. J. 1986. A meta-analysis of age differences in job performance. Journal of Applied Psychology, 71: 33-38.

TABLE 1

Number of Validity Coefficients for Each Test-Criterion Combination Used in This Study

Criteria	General Mental Ability	Verbal Ability	Quantitative Ability	Perceptual Speed	Memory	Spatial/ Mechanical Ability	Motor Ability	Total
Supervisor Ratings	142	277	284	371	73	59	81	1287
Supervisor Rankings	12	19	32	32	8	7	9	119
Production Quantity	22	19	15	39	6	8	18	127
Production Quality	6	16	12	32	7	6	8	87
Work Samples	9	19	15	22	5	7	12	89
Total	191	350	358	496	99	87	128	1709

TABLE 2

Criterion Predictabilities for Seven Test Types

Test Type/ Criterion	Total N	No. rs	Observed Distributions Mean r	Observed Distributions SD	Corrected Prior Distributions $\bar{\rho}$[a]	Corrected Prior Distributions SDρ[b]	Corrected Prior Distributions 90% c.v.
General Mental Ability							
Rating	11,987	142	.25	.155	.44	.197	.19
Ranking	689	12	.40	.232	.66	.334	.23
Work Sample	747	9	.35	.228	.60	.336	.17
Production Quantity	1,116	22	.19	.115	.35	.133	.18
Production Quality	438	6	-.01	.112	-.01	.021[c]	-.04[c]
Verbal Ability							
Rating	24,620	277	.17	.151	.32	.199	.06
Ranking	639	19	.30	.221	.52	.271	.18
Work Sample	1,387	19	.29	.169	.50	.212	.23
Production Quantity	931	19	.16	.130	.28	.023[c]	.25[c]
Production Quality	1,134	16	.08	.131	.15	.106	.01
Quantitative Ability							
Rating	24,913	284	.22	.124	.40	.108	.26
Ranking	1,392	32	.39	.161	.64	.137	.47
Work Sample	1,114	15	.32	.085	.55	.015[c]	.53[c]
Production Quantity	630	15	.25	.133	.44	.023[c]	.41[c]
Production Quality	647	12	.09	.207	.17	.300	-.22
Perceptual Speed							
Rating	30,407	371	.22	.153	.38	.192	.14
Ranking	1,339	32	.25	.161	.45	.101	.32
Work Sample	1,407	22	.36	.088	.61	.015[c]	.59[c]
Production Quantity	1,977	39	.21	.194	.39	.250	.07
Production Quality	2,377	32	.16	.195	.30	.292	-.07

TABLE 2 (continued)

Criterion Predictabilities for Seven Test Types

Test Type/ Criterion	Total N	No. r's	Observed Distributions $\bar{r}$	Observed Distributions SD	Corrected Distributions $\bar{\rho}$[a]	Corrected Distributions SD_ρ[b]	90% c.v.[d]
Memory							
Rating	5,637	73	.17	.142	.32	.161	.11
Ranking	198	8	.19	.173	.35	.031[c]	.31[c]
Work Sample	171	5	.30	.208	.53	.239	.23
Production Quantity	274	6	.21	.089	.38	.016[c]	.36[c]
Production Quality	462	7	.17	.221	.32	.339	-.12
Spatial/Mechanical Ability							
Rating	6,235	59	.13	.123	.24	.139	.06
Ranking	433	7	.21	.193	.37	.267	.03
Work Sample	406	7	.24	.145	.42	.106[c]	.29[c]
Production Quantity	364	8	.17	.181	.30	.197	.05
Production Quality	572	6	.04	.107	.07	.064	-.01
Motor Ability							
Rating	7,860	81	.15	.119	.27	.110	.13
Ranking	302	9	.12	.210	.23	.234	-.07
Work Sample	561	12	.31	.182	.54	.234	.24
Production Quantity	868	18	.23	.203	.42	.263	.08
Production Quality	689	8	.10	.138	.18	.162	-.03

[a]Mean estimated true validity when observed validity is corrected for test unreliability and range restriction.

[b]Estimated standard deviation when variance due to unreliability, rnage restriction, and sampling error have been removed.

[c]In these cases predicted variance was greater than or equal to 100%. Because of the unlikelihood that statistical artifacts actually account for all of the variance in observed validities, these values were recalculated assuming only 90% of the observed variance was due to statistical artifacts. [d]c.v. = credibility value.

TABLE 3

Percentage of Overlap Among Criteria for Both Observed (Below Diagonal) and Corrected (Above Diagonal) Validity Distributions

Criteria	Ratings	Rankings	Work Samples	Quantity	Quality
			General Mental Ability		
Ratings		68	76	79	3
Rankings	70		93	51	5
Work Samples	79	91		59	8
Quantity	85	59	68		1
Quality	33	30	29	45	
			Verbal Ability		
Ratings		67	66	78	45
Rankings	73		97	42	33
Work Samples	71	98		35	27
Quantity	97	69	74		32
Quality	75	53	48	76	
			Quantitative Ability		
Ratings		33	22	76	52
Rankings	55		58	21	28
Work Samples	63	78		0	23
Quantity	91	64	75		40
Quality	69	41	43	64	
			Perceptual Speed		
Ratings		81	27	98	87
Rankings	92		17	86	70
Work Samples	58	66		41	31
Quantity	98	91	59		82
Quality	86	80	48	90	

TABLE 3 (continued)

Criteria	Ratings	Rankings	Work Samples	Quantity	Quality
			Memory		
Ratings		88	60	73	100
Rankings	95		51	52	41
Work Samples	71	77		58	72
Quantity	86	97	76		87
Quality	100	96	76	77	
			Spatial/Mechanical Ability		
Ratings		75	46	83	40
Rankings	80		89	88	36
Work Samples	68	93		69	4
Quantity	90	91	83		38
Quality	70	93	43	65	
			Motor Ability		
Ratings		91	43	69	74
Rankings	93		89	70	90
Work Samples	60	63		81	36
Quantity	80	79	83		58
Quality	85	95	51	70	