WHAT IS THE VALIDITY OF A CONTENT VALID TEST ?

John E. Hunter

Department of Psychology

Michigan State University

March 30, 1981

Presented to International Personnel Management Aoociation,
Minneapolis, June 7, 1982.

Overview

Contemporary legal literature treats a content valid test as if the test had not been shown to have criterion related validity. For example many lawyers have argued that showing content validity does not in itself permit the use of a test for ranking. This can be shown to be logically inconsistent. However it is true that a content validity study does not directly estimate the criterion related validity of the test, it merely asserts that the validity is high.

This study uses the technique of validity generalization to estimate the criterion related validity of a content valid test. This validity turns out to be very high. Perfectly reliable job knowledge and work sample tests correlate .78 on the average with a credibility interval of .70 to .86. If supervisor ratings are used to measure job performance, the validity of a perfectly reliable job knowledge test would be .48 on the average with a credibility interval of .38 to .58.

This level of validity is higher than that associated with ability tests and leads to even greater amounts of economic benefits stemming from optimal use of such tests in selection.

In particular, the validity of the New York police exam is probably .71 and no lower than .64 . This provides a very high basis for economic and administrative advantages stemming from use of the test to rank candidates for selection. That is, this report shows the police exam to have very high criterion related validity and all the benefits that a highly valid test would be expected to have. These benefits will be spelled out in a separate report.

The principles of validity generalization have been accepted by the Federal Court (see Pegues versus State of Mississippi,). However there will be a separate report addressed to the bureaucratic standards in the Uniform Guidelines of 1978 as well as the scientific standards of the American Psychological Association, especially the Division 14 Principles of 1980.

Validity

Scientific jargon constitutes a useful shorthand for communication between experts, but often produces major misunderstandings between experts and other audiences. Nowhere is this more evident than in the legal literature on test validity. Legal experts write as if there three kinds of validity: content validity, criterion-related validity, and construct validity. There is only one kind of validity. There is also misunderstanding about the word "valid" as well.

There is only one meaning to the word "validity" in the area of personnel selection : the validity of a test (or other predictor) is the correlation between test score and job performance computed on the applicant population. A number generated from a validation study should be called an "estimated validity coefficient", though since experts all understand the words "estimated" and "coefficient", they abuse language by calling the number a "validity".

The other word that is abused is "valid". Experts often refer to a test as "valid" or "invalid" as if there were a dichotomy. But in actuality, validity is a quantity not a dichotomous quality; it varies from .00 to 1.00 as the predictive power of the test varies from nil to perfect. The phrase "valid" usually stands for the complex notion "the estimated validity coefficient was statistically significant". A secondary meaning for the phrase "the test is valid" is "the validity of the test is high enough to be useful to the employer". These two meanings are not equivalent.

What is the meaning of "content validity", "criterion related validity", or"construct validity" ? These phrases refer not to kinds of validity but to methods of estimating validity. A test is called "content valid" if a high validity is assured by relating the test content directly to job performance. "Criterion related validity" is established by showing that an estimated validity coefficient from a concurrent or predictive validity study is high enough to be statistically significant(though that is . a debated .. strategy in the profession at the moment. "Construct validity" is the extent to which the score on a test correlates with the theoretical construct which the test is intended to measure. However the term is abused within the personnel selection literature by not distinguishing between the construct validity of the test and the predictive power of the construct that is to be measured. This creates no problems in the field at present since no one has used contruct validity as a way of validating a test. For this reason, construct validity will be ignored in this report.

Criterion related validity

Since validity is the correlation between test scores and job performance on the applicant pool and criterion related validity studies produce a correlation between test scores and job performance scores, it would appear that criterion related validation would be the preferred strategy. This was once the dominant belief in the professional literature, but increased sophistication about the problems with criterion related validity studies has changed that belief.

Criterion related validity studies . differ from the situation required for true validity coefficients in three ways. First, a validation study never measures job performance perfectly. Second, the validation study must be run on an incumbent population rather than an applicant population. Third, the validation study is run with only a small number of workers rather than the full applicant population. Each of these problems leads to error in the estimated validity coefficient. Measurement error and error stemming from the use of incumbent populations are systematic errors and can be reduced by the use of statistical correction formulas. The sampling error stemming from the use of a small sample can neither be estimated nor removed from single studies.

It is very difficult to measure job performance with high precision. Because of limited resources, most studies have been forced to use a very limited measure of job performance : the rating of job performance by the worker's supervisor. A supervisor rating can be deficient in three ways: (1) the rating instrument will not perfectly measure the opinion of the rater, (2) the ratings of different supervisors differ because of leniency differences and rater halo, and (3) raters may not have sufficient opportunity to observe performance.

The extent to which the rating instrument measures the supervisor's impression of the worker is measured by the rating reliability. This will vary from .60 for single rating scales to near 1.00 for composite scores across a set of good rating scales. The extent of leniency and halo is measured by correlating the impressions of two or more supervisors (in those few studies where more than one supervisor knows the worker). Even if the rating scale has perfect reliability, the inter-rater reliability will be surprisingly low. A cumulative study by King, Hunter, and Schmidt(1980) has shown this inter-rater reliability to be about .60 for perfect rating scales. The operational reliability of a given rating scale is the product of the two kinds of reliability. This leads to .60 as an upper bound estimate of the operational reliability of the rating by a single supervisor. Use of .60 as an estimate of operational reliability ~~This~~ leads to a conservative estimate of the validity when it is corrected statistically for error of measurement.

An ideal predictive validity study would hire every applicant until a pool of 1000 applicants had been hired. After a period of time representing full formal and on the job training, job performance would be assessed. If job performance were perfectly measured, then the correlation between job performance and test score would be the validity of the test. If job performance were measured by a supervisor rating, then the observed correlation could be corrected for error of measurement by dividing by the square root of .60 (as noted in the preceding paragraph). But if not everyone is hired, or if less than 1000 workers are used for the correlation, then there are further errors.

The ideal study is rarely done. It is a rare employer who has openings for all applicants. One could evade this problem by hiring a random sample of the applicants, but this makes no sense to the employer. The employer wants to hire the best of the applicants, not a random sample of them. Thus even if the validation study is a predictive study, the people in the final sample will be workers hired on the basis of the test to be validated. Thus the incumbent population will differ from the applicant population in that lower test scores will not be represented. This is called "restriction in range". Restriction in range causes the correlation between test score and job performance score on the incumbent population to be smaller than the correlation would have been for the applicant population. That is, restriction in range causes the estimated validity coefficient to be much smaller than the actual validity.

If the extent of restriction in range is known, then there is a statistical formula that can be used to correct the estimated validity coefficient to obtain an unbiased estimate of the actual validity. This is called "correction for restriction in range."

If the number of workers hired in a given period is small, or if the time period between hiring and completion of full formal and one the job training is long, then a predictive validity study is not feasable and a concurrent validity study will be done. In a concurrent validity study, all current workers on a given job are given the test and are assessed for job performance. If necessary, the correlation between test score and job performance score is corrected for error in measuring job performance. If there are 1000 workers or more in the sample, then the correlation corrected for error of measurement will suffer only from restriction in range. This can be corrected by applying the formula for correction for restriction in range.

Alas it is a rare organization that has as many as 1000 workers employed at a given job. Thus the typical validation study has far fewer workers than are needed for statistical stability. As a result, even after the estimated validity coefficient has been corrected for error in measuring job performance, and even after the correlation has been further corrected for restriction in range, the correlation will still differ from the actual validity coefficient by a random amount called sampling error. The amount of sampling error can never be estimated from the results of a single study. However the potential . size of the sampling error can be computed using a formula called the "standard error". The principal determinant of the size of the sampling error is the number of workers in the study. For 1000 workers, the error bands of the correlation will be about ±.05 and the correlation will be known with 95 percent certainty to one digit accuracy. For fewer workers, the error band will be larger. Lent, Aurbach, and Levin(1971) found that the average sample size for validation studies is only 68. For a study with 68 workers, the error band on the uncorrected correlation would be ±.23 . That is, an observed correlation of .25 would represent an uncertainty interval of .02 to .48 . If the observed correlation must be corrected for error of measurement and for restriction in range, then the endpoints of the uncertainty interval are similarly corrected and the uncertainty interval for the corrected correlation will be still larger than ± .23 .

The only effective way to deal with sampling error is to pool data across studies. This is called "validity generalization within the personnel literature (see for example Schmidt, Hunter, and Pearlman, in press) and is called "meta-analysis" in other areas(Hunter, Note ; Hunter, Schmidt, and Jackson, Note).

The net effect of error of measurement, restriction in range, and sampling error is to render it unlikely that any single organization has the resources to run an effective criterion-related validity study. This basic infeasability of single study criterion-related validity studies was first noted by Schmidt, Hunter, and Urry (1976). Since then psychologists have become convinced that there are only two reasonable methods for criterion-related validity: form a cooperative group across organizations in order to run a basic study with 1000 workers or more, or collect data after the fact from enough studies so that the validity generalization analysis can be done with a cumulative sample size of 1000.

Content validity

The strategy in content validity is to assure that the test will have high validity by controlling the content of the test. The content of the test is related to the nature of the job in such a way as to guarantee the relevance of each item and hence guarantee the relevance of the test to job performance.

Content validation begins with a job analysis; usually a task analysis. In a task analysis, the job is broken down into constituent tasks. These tasks are evaluated for criticality; usually scaled by job experts along dimensions such as importance, consequence of error, frequency, and time spent on task. This information is pooled to determine the critical tasks. This analysis is then used in different ways to construct different tests.

In a work sample test, the tasks are reproduced in the test situation and performance in each task is assessed by observation.

In a job knowledge test, each task is further analyzed in terms of the knowledge required to perform that task. Items are then written to test for that knowledge.

In a prerequisite ability test, each task is further analyzed in terms of the abilities required to perform well at the task. If the abilities used are well known abilities such as verbal comprehension, arithmetic reasoning, etc. ; then test items can be drawn from existing scales. Drawing items from well studied existing scales guarantees that the items will measure the desired ability.

Content validation has none of the problems that plague criterion related validity studies; no error of measurement of job performance, no restriction in range, and above all no sampling error. However content validation has its own problem: there is no numerical estimate of the validity of the test. This has never bothered psychologists because they know that the validity of content valid tests is very high. However it has opened content valid tests to certain strange attacks from lawyers in recent court cases. In particular, lawyers have argued that if a test is "only" content valid, then it is not suitable for ranking. From a scientific point of view, this is illogical. A test has high validity only to the extent that scores on the test correlate highly with job performance. But this correlations is high only if the rank order of persons on the test is very similar to the rank order of persons on job performance. Thus a test can be content valid only if the rank order of persons on the test is highly congruent with the rank order of their level of job performance. That is, a test is content valid only if ranking is a reasonable way to use test scores for selection.

Knowledge and performance

Psychologists have never questioned the validity of content valid tests. Consider a job knowledge test for example. A worker cannot do the right thing unless he either knows or can figure out the right action to take. That is, good performance is predicated on knowledge and the ability to apply that knowledge to ~~in~~ the situation at hand. Thus a high level of job performance implies a high level of job knowledge.

This view has recently been challenged by lawyers. They argue that just because a worker knows the right thing to do, that doesn't mean that the worker will do it. The error in this argument stems from a radical ~~differxxxt~~ difference in ~~profxxxin~~ professional experience. Lawyers deal mostly with behavior in moral terms. To say "Do the right thing" to a lawyer means "Do the morally ~~xxxlxgxllyxxxxxxxxxxt~~ correct or legally mandated thing". A typical behavior choice ~~fxxxxxxxy~~ in the moral domain might be "to take a bribe" versus"report the bribe". The police officer knows the "right" answer, but may be tempted by the money into choosing the "wrong" answer.

Behavior in the work domain is very different from behavior in the moral domain. A typical example of behavior choice would be "Book the suspect under code 1072A" versus "Book the suspect under code 1072B". The choice here is not a matter of morality, but of knowing which category fits the crime and which category does not. Correct behavior will follow from knowledge about the relevant distinction. Or consider an example that requires application of knowledge: "Follow up on Suspect" versus "Temporarily drop investigation of suspect". Here the officer must correctly figure out what knowledge is relevant and must be able to translate that knowledge into the particular facts ~~which~~ in that case. Correct behavior will stem from correct reasoning and adequate knowledge, not from moral considerations.

In the moral domain inhabited by lawyers, "right-wrong" means "good-bad" or "morally sanctioned- morally taboo" or "legally allowed- legally forbidden". In the work domain inhabited by psychologists, "right-wrong" means "correct-incorrect" or "effective-ineffective" or "desired consequences- undesired consequences". In the moral domain there is often a positive reward or temptation for choosing the "wrong" response, and hence there is an element of choice which is independent of knowledge. In the work domain, the reward nearly always goes with the "right" response. A "wrong" choice is called an error or a mistake and usually draws a reprimand. That is, wrong choices in the work domain are usually negatively valued by everyone including the worker under consideration.

One alternative for poor work despite high knowledge would be sloth. The poor worker might choose not to work even though the worker knows what to do. But this is controlled in most work situations by supervision. Few jobs offer the opportunity for undiscovered sloth.

Consider the job of police sergeant. This is a very visible job. Usually the sergeant works in plain sight of many people including other sergeants and especially including his lieutenant and one or more of his subordinates. The sergeant must appear busy. But most people report that if they must do the task, they would rather do it right than do it wrong and have to redo it.

To summarize... Lawyers find it plausible that job knowledge might not be highly correlated with job performance because the erroneously think of "knowledge-behavior" in the moral domain instead of the work domain. Even a brief perusal of the empirical literature on training would reveal the very immediate tie between job knowledge and job performance which has been the staple of the psychologist's experience. In fact, many psychologists with experience in training have argued that job knowledge tests are better measures of job performance than supervisor ratings. Down through the years many studies have used job knowledge tests as criterion measures, i.e as measures of job performance in criterion related validity studies.

<u>The validity of job knowledge tests</u>

This report will not leave the issue of the validity of job knowledge tests at the level of argument. There is a substantial data base for numerically estimating the validity of job knowledge tests. Few of these studies were done with the intention of validating job knowledge tests; psychologists have never doubted the validity of job knowledge tests. However job knowledge tests have occasionally been used as measures of job performance, i.e. as criterion measures in criterion related validity studies. Among these studies are studies which measured not only job knowledge, but other measures of job proficiency as well. That is, there are studies which used work sample tests or supervisor ratings as well as job knowledge tests. These studies permit a numerical measure of the correlation between job knowledge and job performance.

My search for studies was conducted as follows. For several months I have been asking all my friends and asking them to ask friends

Table 1. Validity generalization for job knowledge tests using work sample tests and supervisor ratings as measures of job performance.

Table 1a. Findings for individual studies(See appendix for auth decimals omitted.

Occupation	Sample Size	Validity Work Sample	Validity Supervisor Rating
Cardiologist	1437	78	--
Cardiologist	82	--	60
Customs Inspector	186	79	--
IRS Investigator	292	58	31
Claims Adjustor	175	102	--
Claims Examiner	233	--	63
Electronics Tester	98	--	41
Clinical Lab Technician	160	78	--
Cartographic Technician	443	78	52
Medical Technician	456	--	54
Firefighter	210	80	--
Armor Crewman	368	84	44
Armore Repairman	360	76	34
Supply Specialist	380	83	43
Cook	366	71	64

Table 1b. Validity generalization results using supervisor ratings as the measure of job proficiency.

Average validity	=	.48
Standard Deviation	=	.08
Worst case validity	=	.38
Best case validity	=	.58

Table 1c. Validity generalization results using work sample tests to measure job proficiency.

Average validity	=	.78
Standard Deviation	=	.06
Worst case validity	=	.70
Best case validity	=	.86

for such studies(i.e. in hopes of picking up unpublished studies). At the same time, I searched the published literature back to 1965. Only two relevant studies appeared in the journals(DeNelsky and McKee, 1974 ; Gael and Grant, 1972), and neither published the ~~dxsxrxd~~ desired correlations. However a total of 15 unpublished studies were located. The raw information from these studies is presented in the appendix. The ~~xalidxdxfxx~~ validity generalization results are presented in Table 1.

----------- Insert Table 1 about here --------

The validity of job knowledge tests was assessed against either of two measures of job proficiency: work sample tests or supervisor ratings. A work sample test is a simulation of the job itself with direct observation of performance in that simulation. The supervisor rating ~~wxxxd~~ is usually a composite score of ratings on several dimensions of job performance. Work sample tests have been the preferred measure of performance, because there is evidence that supervisors give more weight than is desired to compliance behavior(i.e. getting along with the supervisor and co-workers, non-deviant dress habits, etc.). However in most research, work samples have not been used because they are very ~~xxpxxxix~~ expensive to run. However, 11 out of the 15 studies in this set did use a work sample test. Therefore analysis could be conducted separately on work sample test validities and supervisor rating validities.

Table 1a presents the individual validity coefficients for each study fully corrected for error of measurement. The validity of job knowledge tests ~~fxxxxxx~~ using supervisor ratings as the measure of job proficiency varies from .31 to ~~.30~~ .64 in terms of

Table 2. Validity estimates for the New York police exam; obtained from Table 1 by setting the reliability of the job knowledge test to .83.

Table 2a. Validity of the police exam using supervisor ratings as the measure of job proficiency.

Expected validity	=	.44
Standard deviation	=	.07
Worst case validity	=	.35
Best case validity	=	.53

Table 2b. Validity of the police exam using work sample tests as the measure of job proficiency.

Expected validity	=	.71
Standard deviation	=	.05
Worst case validity	=	.64
Best case validity	=	.78

Table 3. Validity of job knowledge tests in relation to the educational requirements of the job.

Educational Level Required	Validity using Work Sample Test	Validity using Supervisor Ratings
M.D.	.78	.60
College degree	.76	.45
Techical training	.78	.52
Specific job training	.79	.46

observed correlations. However Table 1b shows that much of this variation is spurious, the product of sampling error. Table 1b shows that had all studies been done with large samples, the credibility interval is .38 to .58 with a typical value of .48. Similarly, the observed values for validity using work sample measures of job proficiency average .78 and vary from .58 to 1.02 (a value that differs from 1.00 by sampling error). Table 1c shows that most of this variation is due to sampling error. The true credibility interval is .70 to .86 with a typical value of .73.

-------- Insert Table 2 about here ------------

The values in Table 1 do not apply to the New York police examination as they stand since they assume perfect measurement on the job knowledge test. The reliability of the police exam is .83 . Table 2 presents the figures for the validity of job knowledge tests corrected to have a reliability of .83. Table 2a shows the credibility interval for validity using supervisor ratings as .35 to .53 with a typical value of .44. Table 2b shows the values of validity using work sample measures of job proficiency to range from .64 to .78 with a typical value of .71.

---------- Insert Table 3 about here ---------

Is there a pattern to predict which jobs have the higher validities? The natural prediction would be that job knowledge tests would be most valid for those jobs which require the most knowledge. The jobs in Table 1 were arranged by educational requirement. The top two studies were both for cardiologists who require an M.D. on top of their college degree. The next four studies are for federal civil service jobs that normally require a college degree. The next four jobs require extensive technical training as a background to the job specific training. The last four jobs require only extensive training specific to the job. Table 3 shows the validity of a job knowledge test for each of these job categories. The validity of the job knowledge test using work sample measures of job performance is essentially constant across categories. The validity of the job knowledge test using supervisor ratings appears to differ for the M.D. category, but a careful check of Table 1a shows the value of .60 is based on a single study with a sample size of 82. For this small sample size, it is quite possible that the deviation of .60 from the average value of .48 is sampling error. Indeed, on this sample, .60 is not statistically significantly different from .48.

The constancy of validity across job categories means that the most likely value for the validity of the police exam is the average value from Table 2b, i.e. a validity of .71. The odds are nine to one against a value as low as .64. Yet even a value of .64 is very high in terms of the implied economic benefit due to the use of the test to select sergeants. A later report would show that benefits of at least one million dollars would flow from a test with a validity even as low as .10.

Conclusion

The cumulative study of the validity of job knowledge tests across occupations shows that the most likely value for the criterion related validity of the New York police examination is .71 . With one chance in 10, the value might be as low as .64; but with one ~~that~~ chance in 10, the valud might be as high as .78. In any case, the police examination has a high ~~degree of~~ degree of validity. ~~[illegible]~~ That is, the fact that the police exam is content valid means that it has a validity of at least .64 in predicting job proficiency. This high degree of validity will be shown to lead to a very high level of economic and administrative benefit stemming from the use of the exam in selecting sergeants for the New York police department.

APPENDIX : RAW CORRELATIONAL DATA FOR THE VALIDITY GENERALIZATION STUDY OF THE VALIDITY OF JOB KNOWLEDGE TESTS

Table A-1 lists the basic results from the 15 studies found which reported the correlation between job knowledge tests and other job proficiency measures. For the most part the numbers are taken directly from the reports. However there were some modifications necessary. In Meskauskas(Note), the two canonical variates for test materials were averaged to obtain an estimate of the sum of the job knowledge tests (thus leading to an average of the canonical correlations). That study was also unusual in that it used multiple raters and hence did not require correction for attenuation in ratings. On the other hand, it used an elite sample for the validation study. The elite sample and reference population standard deviations were both given, so the correlation could be corrected for restriction in range using the usual formula. Thus the shift in correlation from .42 to .54 in this study represents a correction for restriction in range rather than a correction for attenuation due to error of measurement in the job performance measure as is the case in the other studies.

O'Leary(Note) presented data in terms of 5 test scores. However examination of the content showed that tests 2 and 4 were work simulations while tests 1,3,and 5 were job knowledge tests. The data were reanalyzed with the sum of tests 2 and 4 as the work sample test and with the sum of tests 1,3,5 as the job knowledge test.

In three cases, authors ran into construct validity problems. Corts etal(Note) found a very low correlation between ratings by single supervisors. Further investigation showed the supervisors actually had little chance to observe their subordinates at work. Thus they disregarded the supervisor ratings. Schoon etal(Note) found [redacted] supervisor ratings which were not construct valid. Although ratings of the blood banking operation correlated [redacted] .51 [redacted] with job knowledge and .58 with work sample test scores, the ratings for . the other three operations showed negligible correlations. On the other hand, the other three ratings were highly correlated with each other but not with blood banking(despite the fact that job knowledge and work sample scores for all four operations are very high). This strongly suggests that supervisors [redacted] were essentially guessing at levels on the other three operations. Trattner etal(Note) ran into similar problems with their ~~job knowledge~~ work sample test.

-------- Insert Table A-1 about here ---------

The conventional computations of validity generalization are shown in Table A-2. The notation there is of Hunter (Note) and Hunter, Schmidt, and Jackson(Note). Missing reliability coefficients are estimated by the average reliabilities of .81 for job knowledge tests and .84 for work sample tests. Supervisor ratings were assumed to have a reliability of .60 except in the Meskaskas(Note) study as noted earlier.

Table A-1 Basic data for the validity generalization of job knowledge tests using work sample tests or supervisor ratings as measures of job proficiency.

Authors	Occupation	Sample Size
Langdon(Note)	Cardiologist	1437
Meskauskas(Note)	Cardiologist	82
Corts etal(Note)	Customs Inspector	186
O'Leary and Trattner(Note)	Internal Revenue Investigator:or	292
O'Leary(Note)	Social Insurance Claims Adjustor	175
Trattner et al(Note)	Social Insurance Claims Examiner	233
Ramsay(Note)	Electronics Tester	98
Schoon etal(Note)	Clinical Lab Technician	160
Campbell etal(Note)	Cartographic Technician	443
Campbell etal(Note)	Medical Technician	456
Van Rijn and Payne(Note)	Firefighter	210
Vineberg and Taylor(Note)	Armor Crewman	368
Vineberg and Taylor(Note)	Armor Repairman	360
Vineberg and Taylor(Note)	Supply Specialist	380
Vineberg and Taylor(Note)	Cook	366

[a] Corrected for restriction in range, but not for attenuation due to error of measurement in the ratings.

Reliabilities		Raw validity		Corrected for Error in job measure	
Work Sample	Job Knowledge	Work Sample	Supervisor Ratings	Work Sample	Supervisor Ratings
81	--	63	--	70	--
--	--	--	42	--	54[a]
80	72	60	--	67	--
78	64	41	19	46	25
46	66	56	--	83	--
--	81	--	44	--	57
--	72	--	27	--	35
95	91	72	--	74	--
49	88	51	38	73	49
--	85	--	39	--	50
77	78	62	--	71	--
--	81	63	31	76	40
--	76	59	23	66	30
--	92	72	32	80	41
--	84	58	46	65	59

Table A-2. Validity generalization for job knowledge tests.

Table A-2a. Validity generalization for job knowledge tests using supervisor ratings with study by study correctio for error of measurement in both job knowledge test and supervisor ratings

$\bar{\rho} = \bar{r} = .48$

$\sigma_r^2 = .0117$

$\sigma_e^2 = .0047$

$\sigma_\rho^2 = \sigma_r^2 - \sigma_e^2 = .0070$

$\sigma_\rho = .08$

Table A-2b . Validity generalization for job knowledge tests using work sample tests as measures of job proficiency with study by study correction for error of measurement in both ~~[illegible]~~ job knowledge and work sample test.

$\bar{\rho} = \bar{r} = .78$

$\sigma_r^2 = .0060$

$\sigma_e^2 = .0021$

$\sigma_\rho^2 = \sigma_r^2 - \sigma_e^2 = .0039$

$\sigma_\rho = .06$