An Investigation into the Problem of Validity Generalization & Situational Specificity for Nuclear Power Plant Operations

Jeffrey Hornsby
Western Kentucky University

Follow this and additional works at: https://digitalcommons.wku.edu/theses
Part of the Industrial and Organizational Psychology Commons

Recommended Citation
https://digitalcommons.wku.edu/theses/2493
AN INVESTIGATION INTO THE PROBLEM OF
VALIDITY GENERALIZATION AND SITUATIONAL
SPECIFICITY FOR NUCLEAR POWER PLANT OPERATORS

A Thesis
Presented to
the Faculty of the Department of Psychology
Western Kentucky University
Bowling Green, Kentucky

In Partial Fulfillment
of the Requirements for the Degree
Master of Arts

by
Jeffrey S. Hornsby
April, 1983
AUTHORIZATION FOR USE OF THESIS

Permission is hereby

☑ granted to the Western Kentucky University Library to
make, or allow to be made photocopies, microfilm or other
copies of this thesis for appropriate research or scholarly
purposes.

☐ reserved to the author for the making of any copies of this
thesis except for brief sections for research or scholarly
purposes.

Signed [Signature]

Date 03/3/83

Please place an "X" in the appropriate box.

This form will be filed with the original of the thesis and will control
future use of the thesis.
AN INVESTIGATION INTO THE PROBLEM OF
VALIDITY GENERALIZATION AND SITUATIONAL
SPECIFICITY FOR NUCLEAR POWER PLANT OPERATORS

Recommended April 11, 1983
(Date)

[Signatures]

Approved May 3, 1983
(Date)

[Signatures]
DEDICATION

To my family, especially my parents, Rose and Carlo Hornsby. Without their love and support none of this could have been possible.
ACKNOWLEDGEMENTS

I would like to take this opportunity to recognize the contributions made by some extraordinary people. Their help and support made this thesis project a rewarding and enriching effort. I would like to thank my committee: Ray Mendel, John O'Connor, and Jim Craig, for their guidance throughout this research effort. In particular, I would like to thank Ray Mendel. Without his very knowledgeable guidance and unending patience this thesis would not have been possible. I would also like to thank the many friends that I made at WKU--especially, Jeff Prewitt, Marty Cordon, and Craig Sparks. Their help and support helped me persevere to the completion of this project. I would also like to thank Greg Heeter and the rest of the staff at the Center for Nuclear Studies in Memphis, Tennessee, for giving me the opportunity and resources to complete this thesis project.

Finally, I would like to thank the entire Psychology faculty at Western Kentucky University. In addition to my thesis committee, special thanks to Sam McFarland, Sub Fisicaro, Dan Roenker, Leroy Metze, Doris Redfield, and Rich Miller. In one way or another their help and support made my graduate work at WKU a memorable experience.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEDICATION</td>
<td>iii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>vii</td>
</tr>
<tr>
<td>INTRODUCTION AND LITERATURE REVIEW</td>
<td>1</td>
</tr>
<tr>
<td>METHOD</td>
<td>17</td>
</tr>
<tr>
<td>RESULTS</td>
<td>24</td>
</tr>
<tr>
<td>DISCUSSION</td>
<td>30</td>
</tr>
<tr>
<td>REFERENCE NOTES</td>
<td>39</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>40</td>
</tr>
</tbody>
</table>
TABLE OF TABLES

Table 1. Descriptive Data ............................................... 25
Table 2. Results of Situational Specificity Analysis ..................... 26
Table 3. Results of Validity Generalization Analysis .................... 29
Table 4. A Hypothetical Illustration of Validity Generalization When Additional Validity Coefficients are Added to the Results of the Current Study ................. 34
AN INVESTIGATION INTO THE PROBLEM OF
VALIDITY GENERALIZATION AND SITUATIONAL
SPECIFICITY FOR NUCLEAR POWER PLANT OPERATORS

Jeffrey S. Hornsby April, 1983 41 Pages

Directed by: Ray Mendel, John O'Connor, and James Craig

Department of Psychology Western Kentucky University

The issues of situational specificity and validity generalization have shed new light on employment test validation in personnel psychology. Over the past six years Schmidt, Hunter and colleagues have accumulated an abundance of evidence supporting the use of validity generalization. The present study attempts to further investigate the issues of situational specificity and validity generalization for five nuclear power utilities using the CNS test battery. Using the Bayesian statistical procedure developed by Schmidt and Hunter (1977) the results indicated that neither the rejection of situational specificity or the inference of validity generalization was possible. However, further analyses indicated that the failure to reject situational specificity or infer validity generalization was due to one deviant validity coefficient. The addition of further validity studies will increase the likelihood of rejecting situational specificity and inferring validity generalization to a new nuclear power utility without the necessity of another validity study.
Introduction and Literature Review

The belief that test validity is highly situation-specific has long been one of the primary theories in personnel psychology (Schmidt & Hunter, 1977). Variance from study to study in validity coefficients occurs even though job characteristics and situations appear quite similar. The explanation usually given for these results has been that the underlying factor structure of job performance is different from job to job and situation to situation and the job analyst is too poor an information processor to detect subtle differences (Schmidt & Hunter, 1978). This implies that there are undetectable moderator variables that cause validity coefficients to vary across settings. Therefore, validation is required for each situation and the generalization of validity coefficients across settings is impossible. Guion (1976) has stated that this inability to solve the problem of validity generalization is probably the most serious failure of personnel psychology and has restricted the growth of research in the area of personnel selection.

A great deal of attention has been given to this problem over the past six years with the most recent research providing evidence that validity generalization is "technically feasible" in more cases than most researchers realize. Schmidt, Hunter, and Urry (1976) found evidence
which suggests that much of the variance in validity coefficients across situations may be due to statistical artifacts instead of moderating variables. They have found that under typical validation conditions, a valid test will show a statistically significant validity in only approximately fifty percent of the studies. According to Schmidt et al. (1976) this affect can be attributed to any combination of three artifacts that tend to obscure test validities and increase validity coefficient variance across studies.

**Hypothesized Sources of Artifactual Variance**

The first artifact is the differences between studies in criterion reliability. Criterion unreliability attenuates or reduces estimates of validity. Schmidt et al. (1976) cite the example where the true validity of a test is .50 and the job performance measure (criterion) has a reliability of .60. The observed value of the resulting validity coefficient would be \((\sqrt{.60})(.50)\) or only .387.

The second artifact is differences between studies in test (predictor) reliability. Test unreliability has the same attenuation effect on validity estimates as criterion unreliability. However, in this case, test unreliability does not attenuate true validity. Unreliability in the predictor is "real" because in actual selection use the test must be used as is. The goal is to validate or judge the success of the test's predictive ability and to correct for problems in the test itself would be
inappropriate and would result in an overestimation of the test's predictive ability or validity. On the other hand, criterion unreliability is not real because it does not affect the "operational value" of the selection test. According to Schmidt et al. (1976), "We use the test to make predictions of actual job performance, not performance as measured by our imperfect criterion measure" (p. 476). However, even though we should not correct for predictor unreliability, it nevertheless is an artifact insofar as the situational specificity hypothesis is concerned.

The third artifact is differences between studies in range restriction on the predictor. According to Schmidt et al. (1976), if the true validity is .50 and the standard deviation of the applicant group on the predictor is 10 and a validation project is carried out on this test, the actual standard deviation of the validity sample (those actually selected) will be approximately 5. This would result in a validity coefficient of .277 despite a true validity of .50. True validity of a selection test is its validity in the entire applicant group.

The individual effect of each artifact is to increase the necessary sample size to provide a chance of detecting a significant validity when it actually exists, i.e., a reduction in statistical power. Typically, all three artifacts are present so the sample size requirements are multiplied. Schmidt et al. (1976) claim that when this lack of power is combined with what they term "a belief in small
numbers," statistical power is usually drastically reduced resulting in the loss of statistically significant validity coefficients in most testing situations. This belief in small numbers refers to the belief that sample sizes ranging from thirty to sixty are adequate for criterion-related validity studies even though statistical power is very low.

The findings of Schmidt et al. (1976) concerning statistical artifacts have been used to explain most of the variance in validity coefficients across situations for similar job-test combinations. When sample sizes are large enough to give reasonable statistical power and the statistical artifacts mentioned above are corrected, the generalization of validity across situations and jobs should be feasible in most cases.

The main objective of the present study is to further investigate the problems of situational specificity and validity generalization for nuclear power plant operators from five different utilities. The following literature discusses the development of the validity generalization process as it pertains to the investigation of the problems of situational specificity and validity generalization.

Investigation of the Situational Specificity Hypothesis and Validity Generalization

Schmidt et al. (1976) laid the foundation for further investigation of validity generalization by accounting for the statistical artifacts that decreased the feasibility of
such studies. They showed that most of the variance in validity coefficients can be attributed to artificial factors and not the situational specificity hypothesis that most researchers believed. Schmidt and Hunter (1977) followed up their research concerning statistical artifacts with the development of a solution to the problem of validity generalization. The purpose of their study was to further test the hypothesis that validity differences within similar jobs are mostly caused by artifacts and to develop a method of validity generalization using criterion-related validity studies. This method of validity generalization follows this process:

Conceptually, the test of the hypothesis that variation with job-test combinations is essentially zero is relatively straightforward. One need only locate a fairly large number of obtained validity coefficients, convert them to Fisher's z and then subtract from the variance of this distribution variance due to various sources of error...If after subtracting variance due to these various sources, the variance of the distribution of validity coefficients is essentially zero, the hypothesis is confirmed (Schmidt and Hunter, 1977, p. 530).

The sources of error variance considered here refer to the statistical artifacts introduced by Schmidt et al. (1976). Schmidt and Hunter (1977) also identify two new sources of artifactual variance. These are criterion contamination or deficiency and computational or typographical errors. However, since it is difficult to judge their frequency and magnitude, it is currently impossible to statistically correct for them.
Two different methods are cited for developing validity coefficient distributions for validity generalization. The first uses validity coefficients based on the maximum likelihood statistical method used in most validation procedures. This method uses only sample-derived information in the estimation of test validity. The second method is called a Bayesian prior approach and uses relevant validity coefficients accumulated from past studies using the same or similar test-job combinations as well as current sample-derived validity information. Schmidt and Hunter (1977) claim that the latter gives a better estimate of the true validity of the test.

Bayesian Prior Distributions Defined and Applied to Validity Generalization

To obtain a Bayesian prior distribution, sample-derived validity coefficients are combined with a prior distribution of validity coefficients with each coefficient weighted according to its informational value (sample size) to form a posterior distribution. The mean of this posterior distribution is taken as the best estimate of the validity of the test. Credibility intervals (confidence intervals) are then placed around the mean to estimate the level of validity that can be generalized with 95 percent confidence.

Schmidt and Hunter (1977) applied their initial version of the Bayesian analysis to four validity
distributions presented in Ghiselli (1966). Results showed that nearly half of the observed variance in the distributions was accounted for by three statistical artifacts. These statistical artifacts included criterion unreliability, range restriction, and inadequate sample size. After correcting for these artifacts, a conclusion of validity generalization was justified for two of the four distributions.

The initial success of Schmidt et al. (1977) with validity generalization was followed up by Schmidt, Hunter, Pearlman, and Shane (1979). They applied an improved version of their generalization procedure to eleven validity coefficient distributions representing two clerical job families and three validity coefficient distributions representing first line supervisors. This improvement involved two new artifact corrections. The first correction was for between-study differences in test reliability. Although correcting for between-study differences in test reliability contradicts what Schmidt et al. (1976) concluded about correcting for test unreliability, Schmidt et al. (1979) make the case that with this correction the residual variance estimates the effect of true situational differences. The second correction, sampling error, introduces a new artifact not yet considered. Sampling error is believed to substantially increase the variability of validity coefficients from study to study since sample sizes for
each study are less than infinity. Random samples differ in their characteristics, and any statistic (in this case a correlation coefficient) will vary some what from sample to sample. Since the samples used in validity studies are not random, this variability from sample to sample is probably even greater and accounts for a large proportion of the variance in validity coefficients between studies using similar job-test combinations.

Results of the Schmidt et al. (1979) study gave further support for validity generalization in all but two of the distributions analyzed. As a consequence of these results, a new decision rule was introduced. Earlier it was claimed by Schmidt and Hunter (1977) that the situation specificity hypothesis is rejected when, after subtracting variance due to statistical artifacts, the variance is near zero. Schmidt et al. (1979) concluded that such a procedure is "scientifically inappropriate in practice" because it does not take into account the fact that the artifactual variance caused by criterion contamination or deficiency and typographical or computational errors cannot be statistically corrected. Since this artifactual variance does not reflect real situational differences, a more realistic but arbitrary decision rule was adopted. The situational specificity hypothesis should be rejected whenever 75 percent or more of the variance in validity coefficients is accounted for by correcting for the "correctable" statistical artifacts.
Pearlman, Schmidt, and Hunter (1980) added substantially to the evidence regarding situational specificity and validity generalization. Schmidt et al. (1980) improved previous research by adding 21 new distributions of validity coefficients to the 11 previously studied by Schmidt et al. (1979) and by including 24 validity distributions that used training success as criteria instead of proficiency ratings. Schmidt et al. (1980) also investigated the robustness of validity generalization conclusions by reanalyzing the validity coefficient distributions under very conservative assumptions: that sampling error is the only source of artificial variance operating in validity distributions. Pearlman et al. (1980) did this to answer the question of whether they were attributing too much variance to the artifacts and that validity generalization is not as common as they believed. Their results showed that the same four statistical artifacts identified in Schmidt et al. (1979) accounted for an average of 75 percent of the variance in 32 validity coefficients based on job proficiency criteria and an average of 70 percent for the 25 validity coefficients based on measures of success in training. These results led Pearlman et al. (1980) to claim that validity generalization research is useful even when the situational specificity hypothesis is rejected. If after correcting the distribution of validity coefficients for statistical artifacts 90 percent of all values in the
distribution are above a minimum useful level of validity, inferences could be made about the validity of the same test type and job situation without the need of a new validity study. The level of validity at the 90 percent mark represents the lower credibility value for the Bayesian prior. However, the selection of 90 percent was arbitrary in nature. Results for the robustness investigation showed that for proficiency ratings criteria distributions, sampling error alone accounted for an average of 77 percent of the variance predicted on the basis of the four artifacts. Therefore, in these distributions, the three artifacts other than sampling error (criterion unreliability, test unreliability, and range restriction) accounted for only a small amount of the total variance. Pearlman et al. (1980) concluded that this Bayesian procedure is robust and that the question of the procedure overestimating variance due to statistical artifacts is moot.

Schmidt, Gast-Rosenberg, and Hunter (1980) investigated validity generalization results for computer programmers. Schmidt et al. (1980) had two objectives. The first was to present the use of an improved Bayesian prior validity generalization procedure and compare its results to those of the original procedure using the same data. In the original research on validity generalization using the Bayesian prior introduced by Schmidt and Hunter (1977), each of the corrections for criterion
unreliability, test unreliability, and range restriction was computed sequentially, with the others being held constant at their mean. The fact that there is an interaction between these artifacts was ignored. However, Schmidt et al. (1980) claimed this effect was minimal; but to answer some critics of the procedure, a new procedure was developed which calculated the effect of each artifact simultaneously to account for the interaction. The second objective was to add to the evidence demonstrating the "overriding" importance of sampling error, relative to the other artifacts, in the determination of validity generalization. As shown by Pearlman et al. (1980) criterion unreliability, test unreliability, and range restriction only account for a small amount of the variance accounted for by artifacts. Most of the variance was due to sampling error.

Results showed that there was no real difference between the two methods of correcting statistical artifacts for the Bayesian prior procedure. For the four validity distributions investigated, the original procedure accounted for 87 percent of the variance and the new one accounted for 90 percent of the variance. To show the overriding importance of sampling error, the authors reanalyzed the data correcting only for sampling error. No conclusions about validity generalization were changed.

Schmidt, Hunter, and Caplan (1981) investigated the "transsituational generalizability" of the validities of
four types of cognitive tests and a weighted application blank for performance in two petroleum industry job groups. Using the Bayesian prior procedure developed over the past five years, their results showed that the situational specificity hypothesis could be rejected for all four test types as well as the application blank.

After having shown that task differences between similar types of jobs do not affect validity, Schmidt, Hunter, and Pearlman (1981) investigated whether task differences between entirely different kinds of jobs would be important enough to prohibit validity generalization. Using large sample data from Army validity studies, Schmidt et al. (1981) found that the moderating effect of tasks is negligible even when jobs differ greatly in task composition. The authors then concluded that since tasks do not substantially moderate validity, the best candidate for causing variance in validity coefficients is the information processing and problem solving demands on the job. They hypothesized that jobs differing on surface behaviors and tasks may necessitate similar information processing and problem solving demands and, as a result of this, require the same test type to predict success on the job.

Schmidt, Hunter, and colleagues (1977, 1979, 1980) have accumulated an abundance of evidence supporting their Bayesian prior procedure for validity generalization. However, at this time, two major aspects of their research
make their procedure unrealistic in most cases. First, it is not very realistic to find prior distributions of validity coefficients as large as those used in their research. It is nearly impossible to find a prior distribution of approximately 100 validity coefficients for any one position. Second, sample sizes, such as those found in military and clerical positions, are far out of reach in most validity situations. Brown (1981) addressed the first issue when he investigated the situational specificity of the validity of a weighted biographical inventory, the Aptitude Index Battery (AIB), for 12 insurance companies. Using a maximum likelihood method for estimating validity, he found that approximately 62 percent of the observed variance in validity coefficients across companies can be accounted for by statistical artifacts. Based on the 75 percent rule of thumb adopted by Schmidt et al. (1979), the situational specificity hypothesis cannot be rejected. Variance due to situational differences did account for up to 38 percent of the observed variance in validity coefficients. Brown (1981) concluded that this situational variance is inflated due to the inability to correct for two sources of error variance: criterion contamination and typographical errors. However, it was found that 95 percent of the corrected validities were above .188 which gives some evidence for validity generalization. Based on prior experience with insurance groups, Brown (1981) suggested that the 12 insurance
companies could be divided into two groups based on average agent production levels, their use of nonpersonal recruiting sources, and their ability to hire applicants with high AIB scores. He found that nearly 100 percent of the variance in validity coefficients was accounted for within each group.

The fact that Brown (1981) used sample derived information to estimate validity may have increased the residual variance in the validity coefficients. The reason for the increased residual variance in validity coefficients is that sample derived information does not involve as many validity coefficients as the Bayesian prior and therefore may result in a less stable estimates of the tests' validity. However, accounting for 62 percent of variance in validity coefficients across situations is very promising and reinforces the notion that generalization of validity across situations or jobs is more a question of degree than a generalizable/nongeneralizable proposition. Also, Brown (1981) showed the importance of isolating as many of the situational moderators as possible. In this case, dividing the companies into two separate groups based on some hypothesized moderators, such as agent production levels and recruiting sources, allowed nearly all of the variance to be accounted for.

Dunnette (1982) addressed the second aspect that prohibits use of the Bayesian prior approach, small sample sizes. He investigated the issue of validity
generalization across situations for nuclear operators at 72 different nuclear power plants. Since nuclear power plants do not employ hundreds of operators each, sample sizes as low as 24 were utilized. Dunnette (1982) concluded that sample sizes much lower than 24 would cause sampling error to be extremely large. Employing the methodology presented by Hunter, Schmidt, and Jackson (1981) for establishing validity generalization, approximately 63 percent of the variance in validity coefficients was accounted for by sampling error alone. Support was found for validity generalization and the situational specificity hypothesis was rejected for all companies participating in the study.

Brown (1981) and Dunnette (1982) improved the feasibility of validity generalization research by finding positive results in studies that reflect the typical validity situation in most organizations. These results, added to the already accumulated evidence found by Schmidt, Hunter, and colleagues, have led Schmidt, Hunter, and Pearlman (1981) to label the investigation of task differences on test validity a "red herring." Results, using various types of criteria measures, have consistently shown that the moderating effect of tasks on validity coefficients is negligible even when jobs differ greatly in component tasks and is almost nonexistent when task differences are less extreme. Instead of beating the proverbial dead horse, Schmidt et al. (1981) argue that
future validity generalization research should focus on its implications for industrial psychology because it has and will continue to revolutionize personnel selection research.

Object of the Present Study

The object of the present study is to add to the accumulating evidence concerning the method of validity generalization utilized by Dunnette (1982) on the job of nuclear operator for five nuclear power plants (each representing a different public utility). Past research has consistently shown that situational variables do not moderate true validity and validity can be generalized to some extent for most situations using same or similar test-job combinations. This being the case, it is expected that statistical artifacts will account for most of the variance across the five utilities and that the situational specificity hypothesis will be rejected for all companies participating in the study.
Method

Overview

In order to investigate the issues of situational specificity and validity generalization, one begins with a distribution of criterion-related validity coefficients. After obtaining this distribution, each validity coefficient is corrected for criterion unreliability, range restriction, and sampling error. Next, to investigate the question of situation specificity, the variance of these corrected validity coefficients is computed and the variance due to sampling error in the corrected validity coefficients is calculated. If the proportion of variance due to sampling error is at least 75 percent of the variance in the corrected validity distribution, the situational specificity hypothesis is rejected. Finally, to answer the question of validity generalization, a prior distribution is formed by weighting each of the corrected validity coefficients according to its informational value (sample size). The standard deviation of this prior distribution (SDp) is then computed in order to find the credibility value (CV) which identifies the lower limit of validity in which there is 95 percent confidence. If this lower limit of validity is at a useful level, then validity can be generalized. This process, applied to the current...
study, is described in detail below.

Subjects and Utilities Participating in the Criterion-Related Validity Studies

Criterion-related validity studies were carried out for the job of nuclear operator in five nuclear power plants from widely dispersed geographic locations. The nuclear operators involved in each of the five criterion-related validity studies held a high school degree or its equivalent and were at least 18 years of age. For each utility, data on both the total applicant group, as well as the validity sample, were collected in order to perform the necessary computations for restriction of range on the predictor.

Collection of Validity Data

Predictor and criterion scores necessary for criterion-related validity studies were collected in association with the Center for Nuclear Studies (CNS). Predictor scores were obtained from the administration of the CNS test battery used for the selection of nuclear power plant operators. A test battery score, based on a simple linear composite of logical reasoning, numerical ability, arithmetic computation, space relations, reading, and mechanical ability scores was used as the predictor score for each individual in each validity study. Criterion scores were obtained from each utility's nuclear operator fundamental training program which is required by the Nuclear Regulatory Commission (NRC). A final test
average based on a composite of achievement test scores from courses in mechanical physics, heat and thermodynamics, fluid flow, nuclear physics, reactor physics, instrumentation, chemistry, and radiation protection was used as the criterion score for each individual in each validity study. The above test battery and achievement test scores were used to compute an overall criterion-related validity coefficient for each of the five nuclear power plants.

Analysis of Situational Specificity Hypothesis

The situational specificity hypothesis was addressed by using analytic procedures described by Dunnette (1982) and Brown (1981) and also by Schmidt, Hunter, and colleagues (1976, 1977, 1980, 1981). The hypothesis that situational specificity will be rejected when 75 percent of the variance is accounted for by statistical artifacts (Schmidt, Hunter, Pearlman, & Shane, 1979) can be conceptualized as

\[
\frac{\overline{\sigma_e^2}}{\overline{\sigma_{rc}^2}} \geq 75\%
\]

where \( \overline{\sigma_e^2} \) = expected variance due to sampling error across values of \( r \)

\( \overline{\sigma_{rc}^2} \) = expected variance in corrected validity coefficients across utilities.

In order to perform the above calculation, the five
criterion-related validity coefficients were first corrected for criterion unreliability and then range restriction on the predictor. Since the five utilities in this study had no data on criterion reliability, a very conservative estimate of .80 was used for each utility. This is conservative because .80 reflects a high reliability and results in less increase in the corrected validity coefficient than if lower, but probably more realistic, criterion reliability estimates were used (Note 1). The formula used for correcting a validity coefficient for criterion unreliability is

$$r_{xy} = \frac{r_r}{\sqrt{r_{yy}}}$$

where

- $r_{xy}$ = the corrected validity coefficient
- $r_r$ = the raw validity coefficient
- $r_{yy}$ = the reliability of the criterion measure.

The formula for correcting a validity coefficient for range restriction on the predictor is

$$r_{xy} = r'_{xy} \left[ \frac{\sigma_{x}'}{\sigma_{x}} \right] \sqrt{\frac{1-r_{xy}^2+\frac{r_{xy}}{2}\left[ \frac{\sigma_{x}'}{\sigma_{x}} \right]^2}{1-r_{xy}^2+\frac{r_{xy}}{2}}}$$

where

- $r_{xy}$ = validity in unrestricted sample
- $r'_{xy}$ = validity in restricted sample
- $x$ = standard deviation of selection variable in the unrestricted group
\( \bar{x} \) = standard deviation of selection variable in the restricted sample.

After correcting the validity coefficients for criterion unreliability and range restriction, the variance due to sampling error across all of the validity coefficients was calculated. The formula for estimating the variance due to sampling error is

\[
\sigma_e^2 = \bar{c} \left[ (1-r^2)^2 \right] \frac{K}{N}
\]

where \( \sigma_e^2 \) = expected variance due to sampling error across values of \( r \)

\( \bar{c} \) = the mean ratio of the corrected to the uncorrected values of \( r \)

\( K \) = the number of utilities on which validity coefficients have been computed

\( \bar{r} \) = the mean validity across the separate groups

\( N \) = the total sample summed across all utilities.

Finally, the variance in the corrected validity coefficients was computed. If the variance due to sampling error accounted for at least 75 percent of the variance in the corrected validity coefficients, situational specificity would be rejected.

**Analysis of Validity Generalization**

Even if the situational specificity hypothesis cannot be rejected, validity can be generalized if the credibility value is at a useful level of validity. This analysis required three computations which are described below.
First, each corrected validity coefficient was weighted according to its informational value (sample size); and the mean (p) of this new weighted distribution, called the Bayesian prior, was calculated:

\[
\hat{p} = \frac{(r_{c1} X N_1) + (r_{c2} X N_2) + \ldots (r_{cn} X N_n)}{K}
\]

where \(\hat{p}\) = mean of the weighted corrected prior distribution

- \(N\) = sample size for each utility
- \(\bar{N}\) = mean sample size for all utilities
- \(r_{c}\) = corrected validity coefficient for each utility.

Second, the standard deviation of the prior validity coefficient distribution (SDp) was calculated:

\[
\text{SDp} = \sqrt{\frac{\sum r_c^2 - (\bar{r}_c)^2}{N(N-1)}}
\]

where SDp = standard deviation of the prior validity distribution

- \(r_c\) = corrected validity coefficient.

Third, and finally, the credibility value (CV) was calculated to determine above what level 95 percent of the validity coefficients could be expected to fall:

\[
CV = \hat{p} - 1.65 (\text{SDp})
\]
where \( CV \) = credibility value of the prior validity coefficient distribution

\( \hat{\alpha} \) = mean of corrected prior validity coefficient distribution

\( SD\hat{\alpha} \) = standard deviation of the corrected prior distribution

The credibility value reflects the lower limit of the true validity of the particular test type. If this validity is at a useful level, significantly different from zero, validity can be generalized. In other words, if this CV reflects a useful level of validity, a validity study in a similar situation is unnecessary.
Results

A summary of the data used in the analysis of situational specificity and validity generalization is presented in Table 1. Sample sizes, observed validity coefficients, estimated criterion reliabilities, and range restriction data are reported for each participating utility. As seen in Table 1, utility E's observed validity coefficient was not significantly greater than zero (p > .05) and therefore could not be corrected for statistical artifacts. Also, the differences between the unrestricted and restricted means and standard deviations were as expected. The restricted test standard deviation decreased significantly and the restricted mean test performance increased significantly over the unrestricted mean and standard deviation. These results were due to the fact that those in the restricted group (validity sample) were selected on the CNS test battery and therefore those "poor test performers" were not included in the validity analysis because they were not selected and therefore did not participate in the nuclear operator fundamentals training program.

Situational Specificity Hypothesis

Results of the five-utility situational specificity analysis are presented in Table 2.
TABLE 1
DESCRIPTIVE DATA

<table>
<thead>
<tr>
<th>Utility</th>
<th>Sample Size</th>
<th>Observed Validity</th>
<th>Estimated Criterion Reliability</th>
<th>Unrestricted Test SD</th>
<th>Unrestricted Test Mean</th>
<th>Restricted Test SD</th>
<th>Restricted Test Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>45</td>
<td>.27*</td>
<td>.80</td>
<td>63.54</td>
<td>250.59</td>
<td>19.20</td>
<td>328.00</td>
</tr>
<tr>
<td>B</td>
<td>26</td>
<td>.59*</td>
<td>.80</td>
<td>38.84</td>
<td>300.48</td>
<td>19.24</td>
<td>320.54</td>
</tr>
<tr>
<td>C</td>
<td>26</td>
<td>.62*</td>
<td>.80</td>
<td>37.24</td>
<td>271.27</td>
<td>23.62</td>
<td>300.31</td>
</tr>
<tr>
<td>D</td>
<td>40</td>
<td>.58*</td>
<td>.80</td>
<td>35.23</td>
<td>299.53</td>
<td>35.23</td>
<td>299.53</td>
</tr>
<tr>
<td>E</td>
<td>23</td>
<td>-.10</td>
<td>.80</td>
<td>52.43</td>
<td>282.55</td>
<td>18.24</td>
<td>337.70</td>
</tr>
</tbody>
</table>

*Significant at .05 level.

**Utility D's validity study was purely predictive in nature.
### TABLE 2
RESULTS OF SITUATIONAL SPECIFICITY ANALYSIS

<table>
<thead>
<tr>
<th>Utility</th>
<th>Observed Validity</th>
<th>Observed Validity Corrected for:</th>
<th>Criterion Unreliability</th>
<th>Criterion Unreliability and Range Restriction</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>.27</td>
<td>.30</td>
<td>.72</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>.59</td>
<td>.66</td>
<td>.87</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>.62</td>
<td>.69</td>
<td>.83</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>.58</td>
<td>.65</td>
<td>.65</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>-.10</td>
<td>-.10</td>
<td>-.10</td>
<td></td>
</tr>
</tbody>
</table>

**Observed Validities**
- $\overline{r} = .39$
- $\overline{r}^2 = .096$
- $\overline{e}^2 = .023$

**Corrected Validities**
- $\overline{r}_c = .59$
- $\overline{r}_c^2 = .16$
- $\overline{e}_c^2 = .053$
**Observed Validities.** Across all utilities the mean observed validity coefficient ($\bar{r}$) was .39, the variance in observed validity coefficients ($\sqrt{r^2}$) was .096, and the variance due to sampling error ($\sqrt{e^2}$) was .023. However, since there is a large spread in the correlation coefficients across utilities, the obtained correlation coefficients were transformed into Fisher's $Z$ coefficients, averaged, and then transformed back into correlation coefficients. The resulting mean correlation coefficient was .42. Using these statistics, only 24 percent of the variance in observed validity coefficients was attributed to sampling error alone. The remaining proportion of variance (76 percent) may be attributed to other sources of variance--such as differences across studies in range restriction, criterion reliability, other artifacts, or situational specificity.

**Corrected Validities.** Across all utilities the mean corrected validity coefficient ($\bar{r}_C$) was .59, the variance in corrected validity coefficients ($\sqrt{r_C^2}$) was .16, and the variance due to sampling error ($\sqrt{e^2}$) was .053. However, since there is also a large spread in corrected correlation coefficients across utilities, the corrected correlation coefficients were transformed to Fisher's $Z$ coefficients, averaged, and then transformed back into correlation coefficients. The resulting mean correlation coefficient was .67. Using these statistics and correcting all but utility E's validity coefficients for restriction of range
on the predictor and for criterion unreliability, only 33 percent of the variance in corrected validity coefficients was accounted for on the basis of sampling error alone. Therefore, since 75 percent of the variance was not accounted for, the situational specificity hypothesis cannot be rejected in this case.

Validity Generalization

Results for the validity generalization analysis are presented in Table 3. For the five utilities using the CNS test battery for the job of nuclear operator, the corrected prior ($\hat{p}$) was .63, and the standard deviation of the prior ($SD\hat{p}$) was .41. This results in a credibility value (CV) of -.05 at or above which 95 percent of all true validities are expected to lie. Since the credibility value includes a correlation coefficient that is not significantly different from zero, validity cannot be generalized and a validity study in a "new" utility is necessary. However, the corrected prior ($\hat{p}$) of this distribution can be used in a future Bayesian validity study as the best estimate of true validity for that distribution. In this case the the corrected prior ($\hat{p}$) was .63.
TABLE 3
RESULTS OF VALIDITY GENERALIZATION ANALYSIS

<table>
<thead>
<tr>
<th>No. of Prior Validity Distribution</th>
<th>95%CV for Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Job</td>
<td>Test Type</td>
</tr>
<tr>
<td>Nuclear Operator</td>
<td>CNS Test Battery</td>
</tr>
</tbody>
</table>

\( \bar{p} \) = mean of corrected validity coefficient distribution

\( \hat{p} \) = mean of corrected prior validity coefficient distribution

SD\( \hat{p} \) = standard deviation of the corrected prior validity coefficient distribution

CV = credibility value of the prior validity coefficient distribution
Discussion

In the present study, two major questions were addressed. The first, descriptive in nature, was whether or not there were situational moderators that caused validity to be situation-specific across the five utilities participating in the study. As indicated above, the results of the five-utility analysis showed that the three correctable statistical artifacts (criterion unreliability, range restriction, and sampling) accounted for only 33 percent of the variation in observed validity coefficients across utilities. This finding is not consistent with the findings found by Schmidt, Hunter, and colleagues concerning situational specificity. However, a closer inspection of the results in Table 1 shows that most of the variance across utilities is caused by the nonsignificant and highly deviant validity coefficient found or utility E. It is noteworthy that this utility had the greatest predictor range restriction and the highest mean test score of any of the five utilities. This resulted in a validity sample that has very little variability in test scores. This low variability in test scores, coupled with the small sample size available for utility E's validity analysis (smallest of the five utilities), unavoidably resulted in a very unstable estimate of validity, where one or two
deviant pairs of predictor-criterion scores in the sample will significantly alter the resulting correlation coefficient. This result typifies the problem that Schmidt, Hunter, and Urry (1976) termed "a belief in small numbers." Thus in a situational specificity analysis that has but a few validity coefficients, each based on a very small sample size (as is the case in the present analysis), the pattern of resulting coefficients is not unexpected. Based on Schmidt et al.'s (1976) assumption that all tests are valid when there is adequate sample size (true validity coefficients between .35 and .50), as sample sizes increase within each utility's validity sample and as more utilities become available to participate in the situational specificity analysis, the results will reflect the success that Schmidt, Hunter, and colleagues found in the rejection of situational specificity in their research.

Results, such as those found in the present analysis, should encourage further investigation of the situational specificity of the CNS test battery for nuclear power utilities. Schmidt, Hunter, Pearlman, and Shane (1979) emphasized the necessity for further investigation of situational specificity as the first step to the establishment of trait-performance relationships in personnel research:

Application of this model may lead to fairly dramatic progress in the establishment of general principles and theories about trait-performance relationships in the world of work. The first step in the development of general principles and theories in this or any other area is the establishment of stable patterns of relationships
among basic variables. In order to establish such patterns of relationships it is first necessary to demonstrate that the doctrine of situational specificity is false (Schmidt et al., 1979, p. 267).

The second issue, inferential in nature, addressed in this study concerns whether validity can be generalized to a new, but "similar" situation without conducting a validity study. The results of the validity generalization analysis using Schmidt, Hunter, and colleagues' Bayesian procedure are promising with a corrected prior of .63. However, due to the increased variation in the prior distribution caused by utility E's validity coefficient and the small number of utilities participating in the analysis, the credibility value (CV) is -.05. Therefore, since -.05 is not significant from zero, validity can not be generalized to a sixth utility without benefit of a cite specific validity study using the CNS test battery.

In order to illustrate the instability of the Bayesian procedure when so few validity studies are utilized, utility E was removed from the validity generalization analysis. The removal of utility E increased the prior ($\hat{p}$) to .75 and decreased the standard deviation of the prior (SD$\hat{p}$) to .14 resulting in a credibility value of .52 at or above which 95 percent of all true validity coefficients lie. These results show how one deviant correlation coefficient significantly alters the conclusions that can be drawn from an analysis using so few validity studies but they do give promise to further validity generalization
research involving the CNS test battery for nuclear operators. However, in practice one cannot preselect one's data such as the case in the illustration above. The above statistics are given only to illustrate the volatility of validity generalization research when so few validity coefficients are involved.

A further illustration relates the importance of this study to future research concerning the CNS test battery as well as other tests for the electric power industry. Table 4 shows the possible outcomes of the validity generalization analysis when four more significant validity coefficients, just like the ones used in the present study, are added one at a time. Even though the addition of more validity studies would not substantially increase the prior of the distribution, it would stabilize the standard deviation of the prior (SDp) thus allowing an inference of validity generalization to a new test situation. Here again these results are purely hypothetical, but the results do show that the current validity generalization analysis can be used as a building block for continuing research into the generalizability of the CNS test battery.

**Recommendations**

The present study found results consistent with the findings of Schmidt, Hunter, and colleagues regarding statistical power, situational specificity, and validity generalization. As discussed earlier, even though results in the present study do not permit the rejection of the
TABLE 4

A Hypothetical Illustration of Validity Generalization When Additional Validity Coefficients are Added to the Results of the Current Study

<table>
<thead>
<tr>
<th>No. of Validity Coefficients</th>
<th>$r_c$</th>
<th>N</th>
<th>$\hat{p}$</th>
<th>SD$p$</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>.65</td>
<td>40</td>
<td>.6790</td>
<td>.3887</td>
<td>.0376</td>
</tr>
<tr>
<td>7</td>
<td>.83</td>
<td>26</td>
<td>.9815</td>
<td>.3549</td>
<td>.0959</td>
</tr>
<tr>
<td>8</td>
<td>.87</td>
<td>26</td>
<td>.6873</td>
<td>.3289</td>
<td>.1446</td>
</tr>
<tr>
<td>9</td>
<td>.72</td>
<td>45</td>
<td>.7270</td>
<td>.3300</td>
<td>.1825</td>
</tr>
</tbody>
</table>

$r_c$ = corrected validity coefficient

N = sample size of validity sample for the utility

$\hat{p}$ = mean of corrected prior validity coefficient distribution

SD$p$ = standard deviation of the corrected prior validity coefficients distribution

CV = credibility value of the prior validity coefficient distribution
situational specificity hypothesis and the generalization of validity, future research on the CNS test battery in these areas looks promising. In light of the necessity for further research, two recommendations can be made.

First, and foremost, it is necessary that an increased number of utilities participate in such a study. The addition of more utilities to the study would reduce or at least stabilize the standard deviation of the prior, and would, given the accumulating evidence that the CNS test battery is valid, increase the credibility value used to make inferences of validity in a new testing situation (as seen in Table 4). The results of this study show what happens to the credibility value when a nonsignificant validity coefficient is discovered. By increasing the number of significant validity coefficients, the informational value of nonsignificant validities would decrease substantially.

Second, increased sample sizes within the utilities are necessary whenever possible. Increased sample sizes within each utility would serve to reduce the variance due to sampling error and increase the power of detecting validity when it actually exists. Due to the fact that nuclear power utilities only employ a limited number of nuclear operators, it may not be realistic, in the short term, to increase sample sizes. However, in the long run, a long range, longitudinal study using accumulating validity evidence will serve to boost sample sizes within
utilities.

**Implications of Schmidt and Hunter Research**

After finding that most of the variance in observed validity coefficients within single job types can be accounted for by statistical artifacts and that gross differences between job and task makeup produce relatively little variation in test validities, Schmidt, Hunter, and Pearlman (1981) discussed four of the practical implications of the rejection of situational specificity and the inference of validity generalization. Given the strong likelihood that additional utility validity studies will be valid, a review and discussion of these implications is presented.

The first implication is for job analysis in selection research. The accumulating evidence of validity generalization has implications for the type of job analysis needed for personnel research involving aptitude tests. In these situations, according to Schmidt et al. (1981), "fine grained, detailed job analyses tend to create the appearance of large differences between jobs that are not of a practical significance in selection" (p. 175). Instead they advocate the use of broader job analyses that group jobs on the basis of "broad content structure or their similarity in inferred ability requirements-without reference to specific tasks, duties, or behaviors" (p. 175).

The second implication is for criterion construction.
If, as Schmidt et al. (1981) suggest, large differences in tasks do not moderate test validities, then task dimensions within the same job will not moderate test validity either. Therefore, only a measure of overall job performance, as used in the present analysis, is needed for validity and any movement toward the "fractionization" of criteria is unnecessary.

The third implication is for other hypothesized moderators of test validity. The question arises that if task differences between jobs do not moderate validity, does anything? Schmidt et al.'s (1981) belief is that since radical task differences between jobs only slightly moderate validity, other hypothesized moderators are also unlikely to be important. The other moderators usually considered include organizational climate, management philosophy, technology, and applicant pool composition. These types of moderators should show up in the test of the situational specificity hypothesis; but usually after the removal of variance due to statistical artifacts, the remaining amount of variance is too small to allow any operation of "non-trivial" moderator effects.

The fourth, and last implication discussed by Schmidt et al. (1981) is the implication of their procedures for the Uniform Guidelines on Employee Selection Procedures (USEEOC, 1978). These guidelines allow for the use of validity generalization (the Uniform Guidelines call it transportability of validity) when it is demonstrated that
the new jobs, one is trying to generalize to, consist of
the same or substantially the same work behaviors as the
jobs in the original validity studies. Schmidt, Hunter,
and colleagues have supportive evidence that calls for
"broader interpretations" of tasks, duties, or behaviors
instead of a point to point correspondence of the jobs
being analyzed. If aptitude test validities do not differ
significantly when jobs varying widely in tasks, duties, or
behaviors are used, there are very few possibilities that
an aptitude test would be valid in one instance and not
another.

Even though neither the argument for the rejection of
situational specificity or the inference of validity
generalization was possible, it seems that an over zealous
"belief in small numbers" by the researcher was the reason
and the failure to reject the situational specificity
hypothesis or infer validity generalization was not caused
by any real moderator effect. Further investigation of
validity generalization and situational specificity for the
job of nuclear operator will reveal that most of the
restrictions to grouping nuclear related jobs are trivial.
These implications discussed by Schmidt et al. (1981) could
go a long way toward solving the sample size problems that
were realized in the current study. Broader
interpretations of tasks, duties, and behaviors will allow
for the inclusion of many more validity studies in the
analysis of situational specificity and validity generalization.
Reference Notes

References


