Tolerance Intervals: Alternatives to Credibility Intervals in Validity Generalization Research

Roger E. Millsap
Baruch College, City University of New York

In validity generalization research, the estimated mean and variance of the true validity distribution are often used to construct a credibility interval, an interval containing a specified proportion of the true validity distribution. The statistical interpretation of this interval in the literature has varied between Bayesian and classical (frequentist) viewpoints. Credibility intervals are here discussed from the frequentist perspective. These are known as "tolerance intervals" in the statistical literature. Two new methods for constructing a credibility interval are presented. Unlike the current method of constructing the credibility interval, tolerance intervals have known performance characteristics across repeated applications, justifying confidence statements. The new methods may be useful in validity generalization research involving a small or moderate number of validation studies. *Index terms: Bayesian statistics, Credibility intervals, Meta-analysis, Tolerance intervals, True validity distribution, Validity generalization.*

A validity generalization study provides estimates of the mean and variance of true test validities using the results of many individual validation studies. Schmidt and Hunter (1977) proposed that the estimated mean and variance of the true validity distribution be used to construct an interval, centered at the mean true validity, that contains a specified percentage of the distribution of true validities. This interval was denoted a "credibility interval," borrowing the concept from Bayesian statistics (Novick & Jackson, 1974).

Since this early work by Schmidt and Hunter, the calculation of credibility intervals has become a routine step in studies of validity generalization (Callender & Osburn, 1981; Callender, Osburn, Greener, & Ashworth, 1982; Linn, Harnisch, & Dunbar, 1981; Linn & Hastings, 1984; Pearlman, Schmidt, & Hunter, 1980; Schmidt, Hunter, & Caplan, 1981; Schmidt, Hunter, Pearlman, & Shane, 1979). In studies where the variance in true validities remains substantial after removing artifactual variation, the lower limit of the credibility interval is used to justify conclusions about the generalizability of the test's validity across settings and organizations (Osburn, Callender, Greener, & Ashworth, 1983; Pearlman et al., 1980; Schmidt & Hunter, 1981; Schmidt et al., 1981; Schmidt, Hunter, & Pearlman, 1982; Schmidt et al., 1979; Schmidt, Pearlman, Hunter, & Hirsh, 1985).

As described in Schmidt and Hunter (1977), the credibility interval is to have a Bayesian interpretation as an interval centered at the mean of the posterior distribution of true validities. The posterior distribution is the distribution of true validities that is obtained by considering both the investigator's initial beliefs about the distribution of true validities, and the empirical results of many separate validation studies as cumulated in the validity generalization procedure. In Bayesian inference, the initial beliefs of the investigator are for-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 12, No. 1, March 1988, pp. 27–32

0146-6216/88/010027-06$1.55

mally represented in the "prior distribution." These beliefs may simply be hunches, or may reflect previous empirical work. The difference between the prior and posterior distributions indicates how the investigator's initial beliefs have been modified by empirical evidence.

While describing a credibility interval in terms of the posterior distribution, Schmidt and Hunter (1977) stated that validity generalization results provide information about the distribution of true validities that can serve as a prior distribution in future validation research. Further empirical validation is unnecessary if a sufficient portion of this prior distribution lies above an acceptable minimum validity: the lower bound of the credibility interval. Schmidt and Hunter (1977) used the estimated mean and variance of the true validity distribution to construct this interval, assuming a normal distribution for the prior. If the lower bound is acceptably large, Schmidt and Hunter (1977) suggested that further validation studies are unlikely, "when used in a Bayesian analysis, to alter the conclusion that the test is valid" (p. 532). Similar viewpoints are expressed in Schmidt et al. (1979) and Pearlman et al. (1980).

Recent discussions of the credibility interval have omitted explicit reference to Bayesian inference and instead view the interval in frequentist terms, as a predictive indicator for the outcomes of future validation studies. Schmidt et al. (1981) stated that when the credibility interval shows that more than 90% of the estimated true validity distribution lies in the positive range, "positive validity can be expected in new settings and organizations more than 9 times out of 10" (p. 268). Schmidt et al. (1985) suggested that "the process of setting confidence bounds on the true validity is the same process used in all of inferential statistics" (p. 723), and drew an analogy to the confidence interval set by an investigator for the true validity in an individual validation study. This paper discusses the credibility interval from a frequentist perspective.

The Frequentist Viewpoint

In validity generalization research, there is an empirical basis for the prior distribution of true validities. The prior distribution can be regarded as describing an actual "population" of true validities, with a single validation study being a sample of size one from this population. The population of true validities can be defined in various ways, depending on the intent of the investigators. In the employment testing context, the population might be defined as "true validities resulting from all possible validation studies within a particular job family using a given predictor and a given class of criterion measures." More or less specificity can be introduced in this definition. The point is simply that such a population is easy to conceptualize.

The validity generalization procedure uses the results from N validation studies to estimate the mean true validity ρ and the true validity variance σ^2. Let $\hat{\rho}$ and $\hat{\sigma}^2$ denote these estimates respectively. A credibility interval is constructed to give bounds that contain a specified proportion of the true validity distribution. Viewed in frequentist terms, how should such an interval be constructed?

Before constructing the interval, the sense in which the interval will "contain" the specified proportion of the distribution must be clarified. In the frequentist view, the true validity distribution is fixed, and the credibility intervals constructed from sample data will vary due to sampling error. Repeated validity generalization studies, applied to the same true validity population, will generate different credibility intervals. It is impossible to say with certainty whether any specific interval does or does not actually contain the specified portion of the true validity distribution. But it is possible to give a statement concerning the percentage of these intervals which can be expected to cover the specified portion of the distribution. Alternatively, the intervals can be constructed so that they will cover the desired portion "on the average." In either case, the basis for confidence in the intervals lies in a hypothetical set of repeated applications in different samples of validation studies.

In the statistical literature, an interval constructed to contain a specified proportion of the population distribution is known as a tolerance interval (Kendall & Stuart, 1979; Mood, Graybill, & Boes, 1974; Proschan, 1953; Wald & Wolfowitz, 1946; Wilks, 1941). A tolerance interval dif-

fers from a confidence interval in that the former encloses a proportion of the entire population distribution, while the latter is constructed to contain the value of a population parameter. In both cases, the sense in which the intervals "contain" their target values is interpreted with reference to repeated applications in many samples.

Returning to the validity generalization application, assume that the true validity distribution is normal. If ρ and σ^2 are known, a tolerance interval can be constructed to contain a proportion P of the true validity distribution as

$$\rho \pm Z_p\sigma \quad , \tag{1}$$

where Z_p is the standard normal deviate such that $Pr(Z > Z_p) = (1 - P)/2$. In constructing credibility intervals, only the lower bound is of interest. If a proportion R of the true validity distribution is to lie above this bound, the appropriate bound is

$$C_1 = \rho - Z_r\sigma \quad , \tag{2}$$

where $Pr(Z > Z_r) = R$. The interval in Expression 1 and the bound C_1 in Equation 2 are exact in the sense that ρ and σ^2 are known, and the bounds are unaffected by sampling error. It is expected that a proportion R of the future validation studies will have true validities which exceed C_1, assuming that the studies are selected from the population leading to C_1. In the literature, the lower bound of the credibility interval is calculated as C_1.

In practice, ρ and σ^2 are unknown, and the tolerance interval is calculated using the sample estimates $\hat{\rho}$ and $\hat{\sigma}^2$ based upon N validation studies. The constructed interval will now be affected by sampling error. Two types of tolerance intervals exist in the literature for this case, and both are discussed by Proschan (1953). The first type is constructed so that on the average, across repeated applications, the proportion of the true validity distribution covered will be P. This interval was presented by Wilks (1941), and is given as

$$\hat{\rho} \pm t_p(N + 1/N)^{1/2}\hat{\sigma} = \hat{\rho} \pm K_{2p}\hat{\sigma} \quad , \tag{3}$$

where t_p is the value in the Student t distribution such that $Pr(t > t_p) = (1 - P)/2$. The lower tolerance bound above which a proportion R of the true validity distribution will lie is

$$C_2 = \hat{\rho} - K_2\hat{\sigma} \quad , \tag{4}$$

where P is set equal to $2R - 1$ for $R > .50$. Although it can be expected that, on the average, a proportion R of the distribution will lie above C_2, there is no explicit control over how often this will be true across repeated applications. Thus the "confidence level," expressing the degree of expectation that a proportion R of the distribution will lie above C_2, cannot be altered. For this bound, the confidence level is fixed at 50% (Proschan, 1953).

The second type of tolerance interval allows specification of a confidence level γ to be associated with the interval. This interval is expected to cover at least a proportion P of the true validity distribution in γ% of the applications. Wald and Wolfowitz (1946) derived the method of construction for these tolerance intervals. Tables giving the value of K_3 in

$$\hat{\rho} \pm K_3\hat{\sigma} \tag{5}$$

for specified values of P, γ, and N can be found in Bowker (1947), Dixon and Massey (1969), Burington and May (1970), or Beyer (1968). The lower tolerance bound above which a proportion R of the true validity distribution will lie with confidence γ% is found as

$$C_3 = \hat{\rho} - K_3\hat{\sigma} \tag{6}$$

by entering the tables with $P = 2R - 1$ for $R > .50$. The bound C_3 is slightly inaccurate, but adequate for practical use. A more precise bound can be found by consulting tables of one-sided tolerance intervals in Burington and May (1970) or Owen (1958).

Table 1 gives the values of K_2 and K_3 at selected values of N for $P = .80$ and $P = .90$, giving values of $R = .90$ and $R = .95$ respectively. Thus for $R = .95$, 95% of the distribution will lie above the tolerance bounds C_2 and C_3, in the sense discussed earlier. Two different confidence levels ($\gamma = 90\%$, $\gamma = 95\%$) are given for C_3. For comparison purposes, the appropriate value of Z_r in Equation 2 is 1.65 at $R = .95$ and 1.28 at $R = .90$, for all values of N. From the table, it is clear that both K_2 and K_3 are greater than Z_r for all N, with the discrepancy being larger at smaller values of N, as would be expected. The implication is that C_2 and C_3 will always be less than or equal to C_1. The bound C_3

Table 1
Values of K_2 and K_3 at R=.90 and R=.95 for Selected Sample Sizes

	R=.90			R=.95		
		K_3			K_3	
N	K_2	γ=.90	γ=.95	K_2	γ=.90	γ=.95
5	1.68	2.73	3.34	2.34	3.49	4.28
10	1.45	1.98	2.21	1.92	2.54	2.84
15	1.38	1.77	1.93	1.82	2.28	2.48
20	1.36	1.68	1.80	1.77	2.15	2.31
25	1.35	1.62	1.72	1.75	2.08	2.21
30	1.33	1.58	1.67	1.73	2.03	2.14
35	1.33	1.55	1.63	1.71	1.99	2.09
40	1.32	1.52	1.60	1.71	1.96	2.05
45	1.32	1.51	1.57	1.70	1.94	2.02
50	1.31	1.49	1.55	1.69	1.92	2.00
60	1.31	1.47	1.52	1.68	1.89	1.96
70	1.30	1.45	1.50	1.68	1.87	1.93
80	1.30	1.44	1.49	1.67	1.85	1.91
90	1.30	1.43	1.47	1.67	1.83	1.89
100	1.30	1.42	1.46	1.67	1.82	1.87
150	1.29	1.39	1.42	1.66	1.79	1.83
∞	1.28	1.28	1.28	1.65	1.65	1.65

will generally be the lowest of the three bounds, depending upon the level of confidence selected. This is the "price" that must be paid for being able to specify a confidence level for the interval in Expression 5 and the bound in Equation 6.

To illustrate the performance of the three bounds in actual data, Table 2 gives C_1, C_2, and C_3 for a validity generalization study reported by Schmidt et al. (1981). This study examined the generalizability of the validities of four cognitive ability tests for performance criteria in two petroleum industry job groups. The two job groups were classified as "operator" and "maintenance," with job proficiency criteria measured in both cases. The four cognitive tests included mechanical comprehension, chemical comprehension, general intelligence, and arithmetic reasoning measures. The true validity mean and variance estimates in Table 2 are taken from Table 6 (p. 267) in Schmidt et al. (1981). All bounds in Table 2 are calculated for $R = .90$. The bounds C_1 are identical to the credibility interval values given in Schmidt et al. (1981).

The tolerance bounds C_2 and C_3 are uniformly lower than C_1, although C_2 is generally close to C_1. The bound C_3 differs significantly from C_1 in several cases. The bounds for the general intelligence (GI) and arithmetic reasoning (AR) tests are negative, and the mechanical comprehension (MC) test retains a clearly positive bound only for the operator job group. Schmidt et al. (1981) judged test validity to be generalizable if the bound C_1 is positive. If this rule is applied to C_3, only the chemical comprehension (CC) test demonstrates generalizable validity for both jobs. This example illustrates that the tolerance bounds C_2 and C_3 can significantly alter the conclusion of test generalizability when the number of validation studies is low.

Two final points deserve mention. The N to be used in constructing C_2 and C_3 is the number of validity coefficients cumulated in estimating ρ and σ^2, and is unrelated to the sample size within each validation study. Even if the true validity in each study were known, the sample size in the tolerance

Table 2
Estimates of the Mean (ρ) and Standard Deviation (σ) of the True Validity, Credibility Bound C_1, and Tolerance Bounds C_2 and C_3 for MC, CC, GI, and AR Test Data From Schmidt, Hunter, and Caplan (1981)

Test Type/Job Group	N	ρ	σ	C_1	C_2	C_3 γ=.90	C_3 γ=.95
MC/Operator	18	.33	.12	.19	.17	.12	.11
MC/Maintenance	12	.33	.17	.11	.09	.01	-.02
CC/Operator	13	.30	.05	.24	.23	.21	.20
CC/Maintenance	10	.25	.00	.25	.25	.25	.25
GI/Operator	16	.26	.19	.01	.00	-.07	-.10
GI/Maintenance	13	.30	.18	.06	.05	-.03	-.06
AR/Operator	12	.26	.20	.01	-.02	-.11	-.15
AR/Maintenance	11	.15	.16	-.05	-.08	-.16	-.19

interval procedure would be the number of coefficients. However, the researcher is interested in the population of true validities, from which the studies at hand are only a sample.

Secondly, the bounds C_2 and C_3 assume normality of the true validity distribution. There is no known empirical evidence to directly support this assumption. All three bounds C_1, C_2, and C_3 may be inaccurate in skewed distributions. Distribution-free tolerance intervals, which make no assumptions about the form of the population distribution, can be constructed from sample order statistics (Kendall & Stuart, 1979). These order statistics are typically unavailable in validity generalization research, hence this approach will not be pursued here.

Conclusions

When validity generalization research is approached from a frequentist perspective, the true validity distribution is viewed as a potentially realizable "population" of true validities. The validation studies cumulated in the validity generalization procedure are viewed as a sample of size N from this population. The estimates of the mean and variance of the true validity distribution are used to construct a lower tolerance bound, above which a specified proportion of the true validity distribution is expected to lie. Because sample estimates are used to construct the bound, the bound can be expected to vary across validity generalization studies. "Confidence" in any particular bound is grounded in the procedure used to construct the bound, and the expected performance of that procedure across many repeated applications.

From a frequentist perspective, no confidence statements can be attached to C_1 when calculated from sample data. On the average, a proportion R of the true validity distribution will lie above C_2. In other words, for a randomly selected study, it can be stated with 50% confidence that a proportion R of the distribution will lie above C_2 (Proschan, 1953). The bound C_3 allows more stringent confidence statements. Both C_2 and C_3 will generally give lower bounds than C_1, but C_2 and C_3 have known performance characteristics across repeated applications. As N increases, the three bounds will converge to the same value. For small N, the bounds may differ considerably. Validity generalization analyses using fewer than 50 studies are not uncommon (Pearlman et al., 1980; Sackett, Harris, & Orr, 1986). In these cases, investigators may wish to use C_2 or C_3 in preference to C_1.

References

Beyer, W. H. (1968). *Handbook of tables for probability and statistics* (2nd ed.). Cleveland OH: The Chemical Rubber Company.

Bowker, A. H. (1947). Tolerance limits for normal distributions. In Columbia University, Statistical Research Group, *Techniques of statistical analysis*. New York: McGraw-Hill.

Burington, R. S., & May, D. C. (1970). *Handbook of probability and statistics with tables* (2nd ed.). New York: McGraw-Hill.

Callender, J. C., & Osburn, H. G. (1981). Testing the constancy of validity with computer-generated sampling distributions of the multiplicative model variance estimate: Results for petroleum industry validation research. *Journal of Applied Psychology, 66,* 274–281.

Callender, J. C., Osburn, H. G., Greener, J. M., & Ashworth, S. (1982). Multiplicative validity generalization model: Accuracy of estimates as a function of sample size and mean, variance and shape of the distribution of true validities. *Journal of Applied Psychology, 67,* 859–867.

Dixon, W. J., & Massey, F. J., Jr. (1969). *Introduction to statistical analysis*. New York: McGraw-Hill.

Kendall, M., & Stuart, A. (1979). *The advanced theory of statistics, Vol. II*. New York: Macmillan.

Linn, R. L., Harnisch, D. L., & Dunbar, S. B. (1981). Validity generalization and situational specificity: An analysis of the prediction of first-year grades in law school. *Applied Psychological Measurement, 5,* 281–289.

Linn, R. L., & Hastings, C. N. (1984). A meta-analysis of the validity of predictors of performance in law school. *Journal of Educational Measurement, 21,* 245–249.

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the theory of statistics* (3rd ed.). New York: McGraw-Hill.

Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research*. New York: McGraw-Hill.

Osburn, H. G., Callender, J. C., Greener, J. M., & Ashworth, S. (1983). Statistical power of tests of the situational specificity hypothesis in validity generalization studies: A cautionary note. *Journal of Applied Psychology, 68,* 115–122.

Owen, D. B. (1958). *Tables of factors for one sided tolerance limits for a normal distribution* (Monograph SCR-13). Washington DC: Sandia Corporation.

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

Proschan, F. (1953). Confidence and tolerance intervals for the normal distribution. *Journal of the American Statistical Association, 48,* 550–564.

Sackett, P. R., Harris, M. M., & Orr, J. M. (1986). On seeking moderator variables in the meta-analysis of correlational data: A monte carlo investigation of statistical power and resistance to type I error. *Journal of Applied Psychology, 71,* 302–310.

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. (1981). Employment testing: Old theories and new research findings. *American Psychologist, 36,* 1128–1137.

Schmidt, F. L., Hunter, J. E., & Caplan, J. R. (1981). Validity generalization results for two job groups in the petroleum industry. *Journal of Applied Psychology, 66,* 261–273.

Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1982). Progress in validity generalization: Comments on Callender and Osburn and further developments. *Journal of Applied Psychology, 67,* 835–845.

Schmidt, F. L., Hunter, J. E., Pearlman, K., & Shane, G. S. (1979). Further tests of the Schmidt-Hunter validity generalization procedure. *Personnel Psychology, 32,* 257–281.

Schmidt, F. L., Pearlman, K., Hunter, J. E., & Hirsh, H. R. (1985). Forty questions about validity generalization and meta-analysis. *Personnel Psychology, 38,* 697–798.

Wald, A., & Wolfowitz, J. (1946). Tolerance limits for a normal distribution. *Annals of Mathematical Statistics, 17,* 208–215.

Wilks, S. S. (1941). Determination of sample sizes for setting tolerance limits. *Annals of Mathematical Statistics, 12,* 91–96.

Acknowledgments

Preparation of this article was supported in part by PSC-CUNY grant #661191. The author thanks several anonymous reviewers for their helpful comments.

Author's Address

Send requests for reprints or further information to Roger E. Millsap, Department of Psychology, Baruch College, 17 Lexington Avenue, New York NY 10010, U.S.A.