Validity Generalization and Situational Specificity: An Analysis of the Prediction of First-Year Grades in Law School

Robert L. Linn, Delwyn L. Harnisch, and Stephen B. Dunbar
University of Illinois at Urbana-Champaign

Results from 726 validity studies were analyzed to determine the degree of validity generalization of the Law School Admission Test for predicting first-year grades in law school. Four validity generalization procedures were used and their results compared. As much as 70% of the variance in observed validity coefficients could be accounted for by differences in the within-study variability of LSAT scores, simple sampling error, and between-study differences in criterion reliability. The 90% credibility value for the true validities was estimated to be .45, and the average true validity was estimated to be .54. Despite the substantial degree of validity generalization, law school and the year the study was conducted explained significant portions of the residual variance in validities. Thus, some degree of situational specificity of validity remained.

Hundreds of predictive validity studies are conducted each year at academic institutions. Grades for the first semester or year at a college or professional school are correlated with test scores and previous grade average or rank in class. Usually the results for a particular class at a particular institution are viewed in isolation with little, if any, regard for results obtained in other years or at other institutions. Situational specificity of results seems to be implicitly assumed.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 5, No. 3, Summer 1981, pp. 281–289

0146-6216/81/030281-09$1.45

In extreme form, of course, complete specificity of results would destroy the utility of predictive validity studies. To have any practical value, it is at least necessary to assume that the results for one class are applicable to the next class. That is, to be used for admissions decisions, the prediction equation derived from results for one class must be used with a new group of applicants for whom criterion data are unavailable. Thus, some degree of generalizability over time is required. Nevertheless, institutional specificity is still typically adhered to.

In employment settings, the long-standing belief in situational specificity of validity has been strongly challenged in recent years by Schmidt and Hunter and their colleagues (e.g., Pearlman, Schmidt, & Hunter, 1980; Schmidt, Gast-Rosenberg, & Hunter, 1980; Schmidt & Hunter, 1977; Schmidt, Hunter, & Pearlman, in press; Schmidt, Hunter, Pearlman, & Shane, 1979). The belief in the situational specificity of the validity of employment tests is based on "the empirical fact that considerable variability is observed from study to study in raw validity coefficients even when the jobs and tests studied appear to be essentially identical" (Pearlman et al., 1980, p. 373). The small sample sizes that are used in many employment test validity studies clearly have contributed to the large variability in observed validity coefficients. The research program of Schmidt, Hunter, and their col-

leagues has shown that effects of simple sampling error, of criterion unreliability, and of range restriction are sufficient to account for much of the observed variability in observed validities between studies. They have concluded that validity results are quite generalizable across situations. According to Pearlman et al. (1980), "situational specificity is largely an illusion created by statistical artifacts" (p. 399).

It need not be concluded, however, that all of the variability between studies in validities is attributable to statistical artifacts for the idea of validity generalization to be useful. Generalizability may be sufficient to support conclusions that the population validity for a given situation is nonzero, though not necessarily identical to that in other situations. That is, there may be between-situation variability in the population parameters, but that variability may be relatively small and a reasonable credibility interval for it may not include zero.

Partial generalizability may also make it reasonable to base predictions on a weighted combination of the group-specific and the combined-groups regression equation. The results of Novick, Jackson, Thayer, and Cole (1972) and of Rubin (1978) have shown some of the potential value of a Bayesian approach to combining the information about a regression equation from other institutions with the institution-specific information. Rubin (1978), for example, found that empirical Bayes equations predicted slightly better than traditional within-institution least squares equations in cross-validation samples. He also found that empirical Bayes equations were more stable over time than the least squares equations. The latter result is potentially quite important because large shifts in weights make it difficult to explain admissions policies to applicants and other interested parties. This so-called "bouncing-beta problem" is largely the result of statistical artifacts.

In academic settings, there seems little question that the widely used admissions tests have nonzero correlations with first-year or first-semester grades (see, for example, American College Testing Program, 1973; Willingham & Breland, 1977). There is, however, great variation in the magnitude of the observed correlations. The correlations between scores on the Law School Admission Test (LSAT) and first-year grades in law school for the 726 studies analyzed for this paper, for example, ranged from a low of −.08 to a high of .71. Five percent of the correlations were less than .16, while another 5% were .54 or larger. The primary purpose of this paper is to determine the degree to which the variability in the observed validities of the LSAT is attributable to statistical artifacts and to time and institutional specificity.

A secondary purpose is to compare alternative approaches to studying validity generalization. In particular, estimation procedures presented by Callender and Osburn (1980) and those used by Pearlman et al. (1980) were compared. These results were compared with results based on "bare-bones validity generalization" (Schmidt et al., in press), i.e., corrections for sampling error alone. Comparisons were also made with results based on an alternative procedure for estimating the effects of range restriction on variability in observed validities. Finally, estimates of variability due to institutions and due to time were obtained using law school and time period as factors in an analysis of covariance.

Method

Data Source

Schrader (1977) summarized results from 726 validity studies conducted at 150 law schools between 1948 and 1974. Institutional code, year of the validity study, sample size, LSAT standard deviation, and correlation between the LSAT and first-year grade average (FYA) were taken from Schrader's summary. These study results comprised the basic data for all the analyses that are reported below.

Data Analysis

Two general approaches were used to analyze the data. In the first approach, estimates of the variability in validity coefficients attributable to

statistical artifacts were obtained. These estimates were used along with the observed variability and average validity to obtain estimates of residual variation, variation in population parameters, and credibility intervals for population parameters. In the second approach, institutional code and time of study were used as predictors, and estimates of the amount of variability attributable to these factors were obtained.

Eight potential sources of variation in observed validity coefficients are recognized by the basic validity generalization model as outlined by Pearlman et al. (1980):

1. sampling error,
2. differences between studies in criterion reliability,
3. differences between studies in test reliability,
4. differences between studies in degree of range restriction,
5. differences between studies in amount and kind of criterion contamination and deficiency,
6. computational, typographical, and data recording errors,
7. differences in factor structure of tests, and
8. true situational variability. (p. 402)

The approach used by Pearlman et al. (1980) provides estimates of the variation due to the first four statistical artifacts. The sum of these four is the predicted variance, which will be denoted S^2_{pred}. The difference between the variance of the observed validities, S^2_r and S^2_{pred}, is the residual variance, denoted S^2_{res}. Estimates of the mean, M_ρ, and standard deviation SD_ρ of "the distribution of true validities" (i.e., the Bayesian prior distribution; Pearlman et al., 1980, p. 404) are obtained by multiplying the mean and standard deviation by a constant to correct for criterion unreliability and for range restriction. Thus, the distribution of true validities is estimated by adjusting values from the observed validity distribution for criterion unreliability and for range restriction.

In the application in this paper, only the first, second, and fourth statistical artifacts were considered, since only one test, the LSAT, was involved and therefore test reliability was not a source of variability. Variance due to sampling error was estimated using the sample-size weighted average of the estimated sampling error variance of each study (see Pearlman et al., 1980, p. 403). The variance due to differences in range restriction was estimated as outlined in Section C of the Appendix of Pearlman et al. (1980), using 100 as the unrestricted standard deviation of the LSAT. (The standard deviation for all persons taking the LSAT has been slightly larger than 100 at recent administrations.) Since estimates of criterion reliabilities were unavailable for the individual validity studies, it was necessary to use an assumed distribution of criterion reliabilities. The assumed distribution, which is shown in Table 1, is modeled after Pearlman et al. (1980).

The average assumed criterion reliability was .85, which is identical to the empirical estimate of the reliability of the cumulative three-year law school grade-point average reported by Carlson and Werts (1977). The reliability of a cumulative grade-point average might actually be expected to be somewhat larger than that of first-year grades because it is a longer composite. Thus, the .85 average reliability is apt to be somewhat conservative. Indeed, a referee suggested that the Spearman-Brown formula should be used to shrink the .85 for three years to a value of .65 for the first-year average, reasoning that it is one-

Table 1
Assumed Distribution of Criterion Reliabilities for the Validity Generalization Analyses

Reliability	Relative Frequency
.95	.15
.90	.30
.85	.25
.80	.20
.75	.04
.65	.04
.55	.02

third the length. In this paper the more conservative .85 has been used, despite the fact that it is likely to be biased in the direction of underestimating the average true validity because the assumptions of the Spearman-Brown formula are not apt to be satisfied. Grades in law school may be expected to be more homogeneous within a year than across years.

The second set of validity generalization estimates was obtained using the formulas presented by Callender and Osburn (1980). The latter estimates differ in that they are not sample-size weighted. Sample-size weighting could be used with the Callender and Osburn approach, but their approach was used directly without introducing this modification. Rather than using a sum of the variances due to the three statistical artifacts to estimate S^2_{pred}, the latter is estimated by

$$S^2_{pred} = S^2_e + M^2_\rho S^2_a (S^2_c + M^2_c) + M^2_\rho M^2_a S^2_c, \quad [1]$$

where

M's are means and S^2's are variances;
a is the correction for criterion unreliability;
and c is the correction for range restriction.

The estimate of SD_ρ is obtained from the square root of the variance estimate in Equation 10 of Callender and Osburn (1980); see Callender and Osburn (1980) for details of estimating the components going into S^2_{pred} and SD_ρ.

The third set of estimates simply involves the first statistical artifact. That is, between-study differences in criterion reliability and range restriction are ignored.

The final set of validity generalization estimates was stimulated by the observation that the standard procedure for correcting for range restriction tends to result in undercorrection (Linn, Harnisch, & Dunbar, in press). The undercorrection can be explained by the fact that the predictor is not typically the sole basis of selection. Even when the predictor is substantially correlated with the actual basis of selection, the substitution of the predictor for the explicit selection variable in the usual correction for range restriction formula will result in an undercorrection (Linn, 1968; Linn et al., in press). Consequently, the estimates of variance due to differences in range restriction used by the Schmidt and Hunter (1977) and the Callender and Osburn (1980) validity generalization estimates will tend to be too small. To the extent that this is so, the estimates of residual variance and the standard deviation of the true validities will be too large.

The alternative estimate of the variance due to between-study differences in range restriction was based on the empirical relationship of the observed validities and the LSAT standard deviations over the 726 studies. As reported by Linn (in press), there is a strong relationship between the observed validities and the standard deviations. The multiple correlation between the observed validities and the LSAT standard deviation and its square (i.e., the variance) was .58 (Linn, in press). In other words, 34% of the variability in the observed validities was predictable simply from knowledge of the LSAT standard deviation. The variance of the predicted validities based upon the quadratic regression equation was used to estimate the variance due to between-study differences in range restriction in the final approach to validity generalization. Variances due to criterion reliability and sampling error were partialled out in the usual way.

Estimates of the amount of variability due to time and institution were obtained using analysis of variance and covariance procedures. The 27-year period spanned by the validity studies was divided into nine 3-year intervals. Trends in validities and LSAT standard deviations were plotted for the nine time intervals, and tests for trend were performed. Two analyses of covariance were conducted with school and time interval as factors and with LSAT standard deviation and LSAT variance as covariates. Analyses were performed using the observed validities as the dependent variable and repeated using Fisher's Z transformations of the validities. The first ANCOVA included all studies, regardless of number of studies for a given school. Thus, there

were 150 levels of the school factor and 9 levels of the time factor. Several schools had only a single study. The second ANCOVA was limited to schools with four or more studies. This ANCOVA had 106 levels of the school factor and 9 of the time factor and included 640 of the validity studies. Cell sizes were unequal and least squares estimation was used. It was assumed that there was no interaction between school and time.

Results

Validity Generalization

The primary results of the four estimation procedures used in the validity generalization analyses were summarized in Table 2. Except for the differences in the standard deviations of the observed validities and the predicted validities, the results of the Pearlman et al. (1980) and Callender and Osburn (1980) procedures are in close agreement. The differences in those two standard deviations are attributable to the use of sample-size weighted estimates in one case and not in the other. They do not imply a fundamental difference between the procedures. The close agreement between the Pearlman et al. (1980) and the Callender and Osburn (1980) estimates was expected, based on previous results reported by Callender and Osburn, thus supporting the conclusions of Schmidt et al. (in press) that either of these approaches—or a third one due to Schmidt et al. (1979)—can be used with equal confidence.

As expected, the mean estimated true validity from either the Callender and Osburn (1980) or the Pearlman et al. (1980) procedures is substantially above the average observed validity. Indeed, the 90% credibility values are of about the same magnitude as the average observed validity. The credibility values are estimates of the validity above which 90% of the true validities fall

Table 2
Validity Generalization Results

	Estimation Procedure[1]			
Estimate	P, S & H (1980)	C & O (1980)	"Bare-Bones"	Empirical RR
Mean Observed r	.344	.344	.344	.344
Mean Estimated True Validity (M_ρ)	.488	.499	.488	.543
SD Observed r's	.114*	.121	.114*	.121
SD Predicted	.077*	.086	.062*	.101
SD Residual	.085*	.085	.096*	.066
SD Estimated True Validities	.123*	.121	.139*	.072
Percent Variance Accounted for	45.	51.	26.	70.
90% Credibility Value	.330	.344	.310	.451

[1]The estimation procedures are: P, S & H for Pearlman, Schmidt and Hunter, C & O for Callender and Osburn, "Bare-Bones" for use of estimated variance due to sampling error as the only artifact, and Empirical RR for estimates of variance due to differences in range restriction based on variance in validities predicted from LSAT standard deviations and variances.
*Sample size weighted estimate.

and are based on an assumption of a normal Bayesian prior distribution with mean equal to M_p and standard deviation equal to that shown for "estimated true validities" given in Table 2 (see Pearlman et al., 1980, pp. 387-388). Both approaches show that a substantial percentage (45 to 51%) of the variability in observed validities can be accounted for by three statistical artifacts. Considerable variability remains unexplained, however, thus making it reasonable to expect that situational specificity may contribute some of the observed variation.

The "bare-bones" estimates indicate that about one-fourth of the observed variability may reasonably be attributed to simple sampling error. Although substantial, this value is smaller than has been found in some of the work in employment settings. The smaller percentage accounted for by sampling error is partially attributable to the relatively good size samples typically used for the law school validity studies (the sample size for the 726 studies ranged from 31 to 1,117 with an average N of 226 and a standard deviation equal to 154), partially to greater variation due to between-study differences in range restriction and partially to greater variance due to situational specificity. The results in the last column of Table 2 provide an indication of the amount of additional variance potentially attributable to between-study differences in range restriction. Situational specificity effects are discussed in the following section.

The results in the last column of Table 2 show a higher estimated true validity, a larger percentage of variance accounted for, and a smaller standard deviation of estimated true validities than either the Callender and Osburn (1980) or the Pearlman et al. (1980) procedures. These differences are attributable to the undercorrection that is obtained from the usual range restriction correction formula in situations that hold for law schools, where the LSAT is an important component of the selection decision but is not the sole determining factor. The mean estimated true validity of .543 is simply the quadratic regression-based estimate of validity, where the test standard deviation is 100, divided by the square root of the mean criterion reliability of .85.

Using the standard formula to correct for range restriction (assuming explicit selection of the LSAT) results in an estimated variance attributable to between-study differences in range restriction of .0012 (this is the S_c^2 in Equation 1 and in the Callender and Osburn, 1980, notation). That is, the standard correction would account for only 8% of the variance in observed validities. The variance in validities predicted from the regression on LSAT standard deviation and LSAT variance, on the other hand, is .0049. Thirty-four percent, an amount greater than that attributable to sampling error, can be accounted for by the LSAT standard deviations and variances in the study samples.

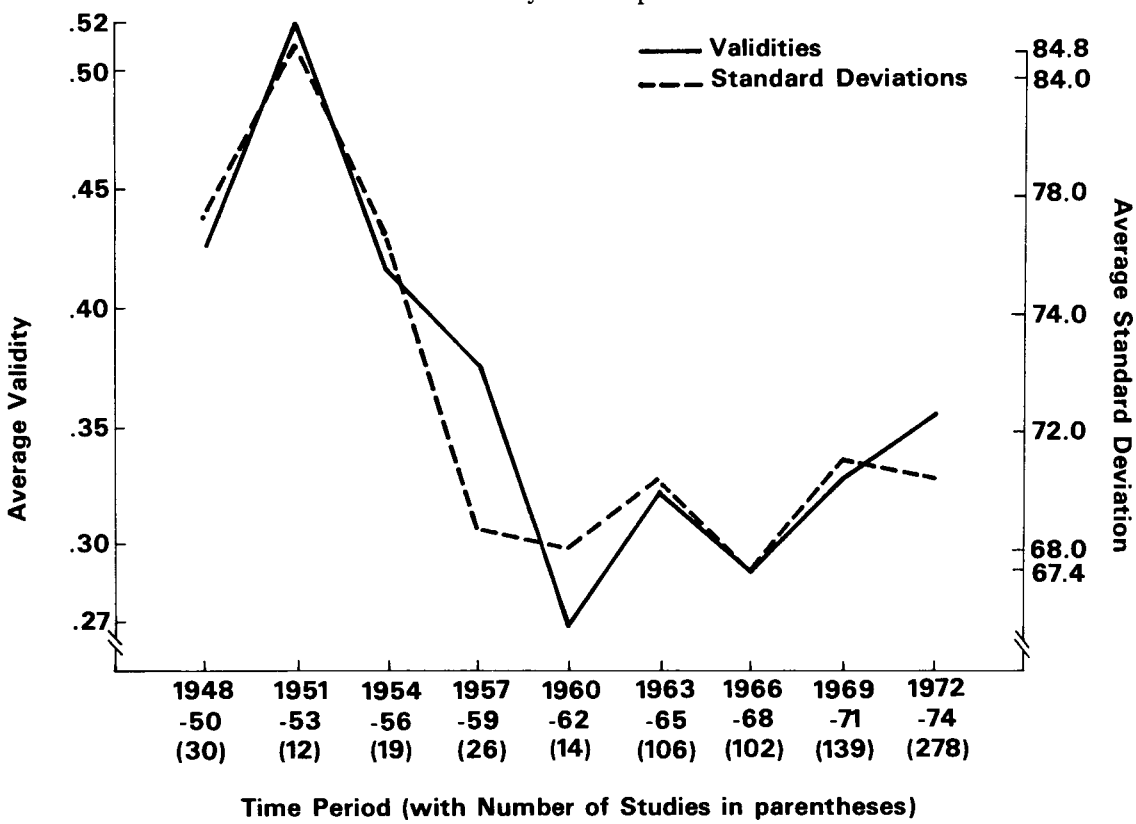
Situational Specificity Effect

Significant differences ($p < .01$) in average validity were obtained when the studies were classified into nine 3-year time intervals starting with 1948 to 1950. The average validities for each 3-year time period are plotted in Figure 1. As can be seen, the average validity declined from the early time periods, reaching a low in 1961 to 1963 and then increased slightly in the more recent time periods. The linear, quadratic, and cubic trend components are all significant at the .01 level.

The dashed line in Figure 1 shows the average standard deviation (see right-hand scale) for each 3-year time interval. As might be expected, there is a substantial similarity between the average standard deviation and average validity plots. The correlation between the two sets of averages is .93.

The results of the two analyses of covariance with Fisher's Z transformation of study validity as the dependent variable, law school and time interval as the independent variables, and LSAT standard deviation and variance as the covariates are summarized in Table 3. Analyses using observed validity as the dependent variable were also run but are not reported here, since they yielded results nearly identical to those using

Figure 1
Plot of average validity (solid line) and average standard deviation (dashed line) for nine 3-year time periods



Fisher's Z transformations. The results in the first half of Table 3 include all 150 law schools, regardless of the number of studies per school. Those in the second half of the table are based on the 106 schools with at least four studies. In both analyses it was assumed that there was no interaction between law school and time interval.

The results of the analysis with all 150 law schools and of the analysis with the 106 law schools that contributed four or more studies are quite similar. In both analyses there were significant main effects for law school and time interval. Even after adjusting for differences in LSAT variability, systematic differences remained due to time interval and law school, suggesting that situational specificity cannot be completely ignored. For example, after adjusting for LSAT variability, the average validity of the 11 studies conducted in various years at one law school was .13 higher than expected, whereas the corresponding figure based on 6 studies conducted at another law school was .13 lower than expected.

Discussion

There is a strong evidence of validity generalization for the LSAT in predicting first-year grades in law school. As much as 70% of the variance in observed validities could be predicted from knowledge of the LSAT variability within a study, simple sampling error, and assumed differences between studies in criterion

Table 3
Analyses of Covariance Summary Results for Z Transformed Validity Coefficients with LSAT Standard Deviation and Variance as the Covariates

Source	df	Mean Square	F[1]
All 150 Law Schools and 726 Studies			
Main Effects			
Law School	149	.029	2.75
Time Period	8	.025	5.40
Residual	566	.009	
The 106 Law Schools with 4 or more Studies per School (Total of 640 Studies)			
Main Effects			
Law School	105	.030	3.16
Time Period	8	.049	5.18
Residual	524	.009	

[1]All F ratios are significant at the .05 level or less.

reliability. Simple sampling variability alone could account for 26% of the variance of the observed validities and another 34% could be predicted from observed standard deviations and variances on the LSAT. Nonetheless, systematic differences were also found between law schools and between time intervals during which the study was conducted. In other words, there was evidence of some situational specificity; not all of the variability could be attributed to statistical artifacts.

The generalizability of validity across law schools is more than adequate to support the conclusion that the true validity is nonzero without the need for a situation-specific study. The 90% credibility value for the fourth validity generalization procedure was .45. Even if .995 were used, the credibility value would still be .36. Thus, the focus on situational specificity is only concerned with factors that influence the magnitude of the validity, not ones that destroy or make the validity too small to be useful. Still, it would be desirable to know the extent to which the curricula and grading practices of law schools that consistently have higher (or lower) validities than would be expected on the basis of the variability in LSAT scores affect the kind of situational specificity observed in this analysis. Changes in grading practices over time could affect both the mean level of grades and their reliability. There could be consistent differences among schools in reliability of grades. No evidence was available regarding such possibilities, but they could contribute to the apparent situational specificity in validity that was observed in the analyses of covariance.

There is a degree of uncertainty in all of the validity generalization methods used. All of the methods require simplifying assumptions. For example, the range restriction correction used in the Callender and Osburn (1980) and in the Pearlman et al. (1980) analyses adjusts for range restriction as if it were due to univariate selection on the basis of the LSAT and relies on standard linearity and homoscedasticity assumptions. Consequently, some caution in interpretation is called for. It seems clear, however, that although the precise degree of generalizability and specificity of validities may be debatable, a substantial fraction of the variation in validities may be attributed to statistical artifacts—in particular, to sampling variability, to variation in range restriction and to criterion unreliability. It is also evident that focusing on

either the mean or the lower tail of the distribution of observed validity greatly understates the predictive value of a test such as the LSAT.

References

American College Testing Program. *Assessing students on the way to college: Technical report for the ACT assessment program, Vol. 1.* Iowa City: American College Testing Program, 1973.

Callender, J. C., & Osburn, H. G. Development and test of a new model for validity generalization. *Journal of Applied Psychology,* 1980, *65,* 543-558.

Carlson, A. B., & Werts, C. E. Relationships among law school predictors, law school performance, and bar examination results (Report No. LSAC-76-1). In Law School Admission Council, *Reports of LSAC-sponsored research: 1975-1977* (Vol. 3). Princeton NJ: Law School Admission Council, 1977.

Linn, R. L. Admissions testing on trial. *American Psychologist,* in press.

Linn, R. L. Range restriction problems in the use of self-selected groups for test validation. *Psychological Bulletin,* 1968, *69,* 69-73.

Linn, R. L., Harnisch, D. L., & Dunbar, S. B. "Corrections" for range restriction: An empirical investigation of conditions resulting in conservative corrections. *Journal of Applied Psychology,* in press.

Novick, M. R., Jackson, P. H., Thayer, D. T., & Cole, N. S. Estimating multiple regression in *m* groups: A cross validation study. *The British Journal of Mathemical and Statistical Psychology,* 1972, *25,* 33-50.

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

Rubin, D. B. *Using empirical Bayes techniques in the law school validity studies.* (Report No. LSAC 78-1a). Law School Admission Research, Princeton NJ: Law School Admission Council, 1978.

Schmidt, F. L., Gast-Rosenberg, I., & Hunter, J. E. Test of a new model of validity generalization: Results for computer programmers. *Journal of Applied Psychology,* 1980, *65,* 643-651.

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

Schmidt, F. L., Hunter, J. E., & Pearlman, K. Progress in validity generalization: Comments on Callender and Osburn and further developments. *Journal of Applied Psychology,* in press.

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

Schrader, W. B. Summary of law school validity studies, 1948-1975 (Report No. LSAC-76-8). In Law School Admission Council, *Reports of LSAC-sponsored research: 1975-1977* (Vol. 3). Princeton NJ: Law School Admission Council, 1977.

Willingham, W. W., & Breland, H. M. The status of selective admissions. In Council on Policy Studies in Higher Education, *Selective admissions in higher education.* San Francisco: Jossey-Bass, 1977, 66-244.

Acknowledgments

We thank Frank L. Schmidt, John E. Hunter, Kenneth Pearlman, and two unknown referees for helpful comments on an earlier version of the paper.

Author's Address

Send requests for reprints or further information to Robert L. Linn, Department of Educational Psychology, University of Illinois, 1310 South Sixth Street, Champaign IL 61820.