DOCUMENT RESUME

ED 281 865	TM 870 249
AUTHOR	Boldt, Robert F.
TITLE	Generalization of GRE General Test Validity across Departments.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	Graduate Record Examinations Board, Princeton, N.J.
REPORT NO	ETS-RR-86-46; GREB-82-13P
PUB DATE	Dec 86
NOTE	27p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*College Entrance Examinations; Departments; Grade Point Average; *Graduate Study; Higher Education; Hypothesis Testing; *Predictive Validity; Sample Size; *Test Theory; *Test Validity
IDENTIFIERS	*Graduate Record Examinations; GRE Validity Study Service; Range Restriction; *Validity Generalization

ABSTRACT

This study of the validity of the Graduate Record Examinations (GRE) General Test used data from predictive validity studies that were conducted by the GRE Validity Study Service (VSS) in 79 graduate departments. The performance criterion was first-year grades in graduate school. Observed validities were computed, and for each graduate department validities were also estimated for groups at two other stages of selection--applicants for admission to the department, and all GRE takers. Two hypotheses were tested: (1) General Test's validities were equal across studies; and (2) General Test's validities had equal ratios across studies, i.e., the level of validities might vary from institution to institution, but the ratios would be constant. These hypotheses were applied for VSS groups, applicant groups, and all GRE takers, and implied validities were calculated. When the implied validities were compared to the observed validities, it was found that the assumption of equal validity did not account well for differences in the level of observed validity of the GRE General Test. The equal ratio hypothesis accounted for the observed validities rather well, but departmental discipline was not significantly related to the degree of fit of observed to implied validities. At all levels of selection, the study yielded applicant validities that were predominantly positive. This lends support to the presumption that the General Test's validity is transportable, i.e., institutions that do not use the General Test can, if they adopt it, expect it to prove valid. Appendices include: (1) use of test theory to present the effects of self selection; (2) use of a supplementary variable when data are missing for an explicit selector; (3) generalizing the assumption that the validities are proportional across institutions; and (4) calculating validities in the restricted group. (Author/JAZ)

GRADUATE RECORD EXAMINATIONS

GENERALIZATION OF GRE GENERAL TEST VALIDITY ACROSS DEPARTMENTS

Robert F. Boldt

GRE Board Professional Report No. 82-13P
ETS Research Report 86-46

December 1986

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

ETS

EDUCATIONAL TESTING SERVICE, PRINCETON, NJ

GENERALIZATION OF GRE GENERAL TEST VALIDITY ACROSS DEPARTMENTS

Robert F. Boldt

GRE Board Professional Report No. 82-13P

December 1986

GENERALIZATION OF GRE GENERAL TEST VALIDITY ACROSS DEPARTMENTS

Abstract

This study of the validity of the GRE General Test used data from predictive validity studies that were conducted by the GRE Validity Study Service (VSS) in 79 graduate departments. The performance criterion was first-year grades in graduate school. Observed validities were computed, and for each graduate department validities were also estimated for groups at two other stages of selection--applicants for admission to the department, and all GRE takers.

Two validity generalization hypotheses were tested. One was that the General Test's validities were equal across studies; the other was that the General Test's validities had equal ratios across studies, that is, that the level of the validities might vary from institution to institution but the ratios would be constant. These hypotheses were applied for VSS groups, applicant groups, and all GRE takers, and implied validities (validities that would be observed if the hypotheses were true) were calculated. When the implied validities were compared to the observed validities, it was found that the assumption of equal validity did not account well for differences in the level of observed validity of the GRE General Test. The equal ratio hypothesis accounted for the observed validities rather well, possibly due to overcapitalization on chance, but departmental discipline was not significantly related to the degree of fit of observed to implied validities.

At all levels of selection, the study yielded applicant validities that were predominantly positive. This lends support to the presumption that the General Test's validity is transportable, i.e., institutions that do not use the General Test can, if they adopt it, expect it to prove valid. In view of the scarcity of very low or negative validities, studies revealing such validities should be questioned.

GENERALIZATION OF GRE GENERAL TEST VALIDITY ACROSS DEPARTMENTS

R. F. Boldt

Traditionally, research on admissions testing has emphasized the results of local validity studies, i.e., separate studies usirg only data from individual institutions, without regard to data collected at other, possibly similar, institutions. This practice, reinforced by the variation in test validities from institution to institution, has been regarded as consistent with professional standards for test use, which have embraced the notion that success may indeed be more predictable at some institutions than at others. These beliefs were also widely held in industrial applications of testing, in which test validity was thought to be highly specific to particular situations. For example, as late as 1975 the American Psychological Association's Division 14 (Industrial and Organizational Psychology) stated in its Principles for the Validation and Use of Personnel Selection Procedures that:

> Validity coefficients are obtained in specific situations. They apply only to those situations. A situation is defined by the characteristics of the samples of people, of settings, or criteria, etc. Careful job and situational analyses are needed to determine whether characteristics of the site of the original research and those of other sites are sufficiently similar to make the inference of generalizability reasonable. (p.13)

An even more extreme view was espoused by the Equal Employment Opportunity Commission's (EEOC) Guidelines on Employee Selection Procedures (1978), which required every use of an employment test to be validated.

However, research on institutional differences in test validity (Schmidt and Hunter, 1977; Schmidt, Hunter, Pearlman, and Shane, 1979; Pearlman, Schmidt, and Hunter, 1980; Schmidt, Gast-Rosenberg, and Hunter, 1980) led increasingly to awareness that the effects of numerous presumed-to-be important variables were far smaller than supposed. In fact, much of the observed variation in test validity could be explained by statistical artifacts, most notably error resulting from the use of small samples and differences among institutions in (a) the effects of selection on the distribution of test scores and (b) the reliability of the criterion. This growing awareness was reflected in the 1980 version of the American Psychological Association's Division 14 Principles, as follows:

> Classic psychometric teaching has long held that validity is specific to the research study and that inability to

> generalize is one of the most serious shortcomings of selection psychology (Guion, 1976). [But]...current research is showing that the differential effects of numerous variables may not be as great as heretofore assumed. To these findings are being added theoretical formulations, buttressed by empirical data, which propose that much of the difference in observed outcomes of validation research is due to statistical artifacts... Continued evidence in this direction should enable further extensions of validity generalization. (p.16)

In addition to acceptance by Division 14 of validity evidence from generalization studies in the personnel sphere, more general acceptance has been won. The American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) have approved a revised edition of Standards for Educational and Psychological Testing (AERA, APA, NCME, 1985). Principle 1.16 in these Standards states:

> When adequate local validation evidence is not available, criterion-related evidence of validity for a specified test use may be based on validity generalization from a set of prior studies, provided that the specified test-use situation can be considered to have been drawn from the same population of situations on which validity generalization was conducted.

This increased acceptance of results of prior studies as evidence of validity at a new site or institution, called validity generalization, is quite welcome because the GRE Program does indeed have difficulties in conducting local validity studies. Quite often the number of cases available from an institution is small. Validity generalization offers possible relief from this problem because, in this approach, the number of cases is increased through the pooling of data from many institutions, and the simplifying assumption about the relationships among validities reduces the number of parameters to be estimated.

Three types of approaches to validity generalization have been discussed in the literature. Schmidt and Hunter, using data gleaned from the published literature, used single selector range restriction theory (Gulliksen, 1950, chap. 11; Thorndike, 1982 pp.208-212) and classical test theory (Gulliksen, 1950 chap. 3; Lord and Novick, 1968 part 2). The estimation procedures available to these authors were limited because various useful data, especially applicant statistics and test reliabilities, were not available. Schmidt and Hunter's type of study is most usually associated with the term "validity generalization," but their analysis was not used for this study because applicant pool data and test reliabilities were available.

A second approach to validity generalization is that utilized by Linn and Hastings (1984). They examined law school admissions data obtained from the Law School Admission Test (LSAT) program, for which applicant pool data were available. The approach taken by Linn and Hastings featured regression analyses with the law school as the data point. They developed regressions of LSAT validities on other statistics, such as the standard deviation of LSAT scores and the correlation between LSAT scores and the undergraduate grade point average for admitted students. This procedure has been used by others--for example Baird (1983)--but has not usually been referred to as validity generalization research. It pertains to the transportability aspect of validity generalization since regression formulas could be used to estimate the validity of the LSAT for a school newly adopting the test, and the correlation coefficient could indicate the precision of the estimated validity. Since this procedure uses multiple regressions and entails the validation of various combinations of variables in the selection of most valid predictor combinations, a very large number of validity studies are needed to avoid excessive capitalization on chance; otherwise substantial overestimates of the amount of validity generalization can result.

A third approach is to use test theory and range restriction theory--Schmidt and Hunter--but to use the multivariate version of range restriction theory together with applicant data from the institutions (sites) studied and data from the total pool of examinees. This approach was available to Linn and Hastings (1984), but they chose not to use it, citing a standard proposed by Lord and Novick (1968, p. 147). The standard cited is that when the ratio of the test score standard deviation in the applicant pool to that in the admitted pool exceeds 1.4, the use of range restriction assumptions is questionable. A ratio of 2.0 is to be regarded as extreme. Of 154 schools in the Linn and Hastings study the ratio of 1.4 was exceeded by 45; for three of these the ratio exceeded 2.0. With such data it is reasonable to seek another approach, such as the second one mentioned above, instead of staying close to the test theory and range restriction approaches of Schmidt and Hunter.

In graduate education, selection is apparently not as extreme as in the law school context. The numbers of departments in the current study for which the ratio of the standard deviation of applicant group scores to those of the VSS group exceeded 1.4 were 6, 5, and 11 for the verbal, quantitative, and analytical measures, respectively, distributed over 16 departments. In no case did the ratio equal or exceed 2.0, and in most cases the ratios that exceeded 1.4 did not exceed it by much. Thus, the degree of selectivity of the admissions process does not preclude the use of the range restriction model. The present study uses this third approach to validity generalization.

Despite these differences, the present study and other validity generalization studies have a common focus on the distribution of validities after statistical artifacts are removed. The study reported here considered four preadmission variables: the verbal, quantitative, and analytical scores from the GRE General Test, and undergraduate grade point average (GPA).

Generalization Hypotheses

Because validity generalization research is concerned in part with the effects of selection on apparent test validity, the present study considered groups of examinees at three stages of selection. The first level was that of "all test takers," i.e., those who took the GRE General Test during a given period of time. This group served as a standard population on which selection had not yet operated. The second level was that of "applicant pools," which consisted of General Test examinees who had applied to one sample of graduate departments included in the study, and who differed from "all test takers" by virtue of the effect on true test score distributions of various social forces that influence application behaviors. (The effect of these forces on the distribution of test scores is discussed in Appendix A.) The third level consisted of examinees who were admitted to departments for whom validity studies had been conducted (Graduate Record Examinations Board, 1985). Having been (a) previously sorted to applicant groups by self selection, (b) selected by departments on the basis of scores on the GRE General Test and undergraduate record, and (c) persistent in completing the first year of college, these "VSS groups" were therefore the most highly selected of those considered here.

This study tested the following generalization hypotheses in groups at each of the three levels of selection mentioned above:

1. that validities for a measure were the same for all institutions, and

2. that, although the validities were not the same for all institutions, the ratios of the scores' validities (verbal to quantitative and quantitative to analytical) were the same.

The second hypothesis allowed for variation in criterion reliability among departments.

Procedures

Samples

Two convenient sources of GRE data were available: test analyses and the student history file. Test Analyses contain statistical data describing examinees from particular administrations. For this project, data from General Test forms 3DGR1, 2, and 3, which were administered in 1981, were combined to provide estimates of GRE standard deviations, intercorrelations, and reliabilities for the group referred to as "all GRE takers." The student history file contains data on all examinees, including those who ultimately attended departments that conducted validity studies. General Test scores and undergraduate GPAs were available, as were responses to the background information questions (BIQ) on the General Test registraion form. For a student's data to be included as part of a VSS group in this study, a complete set of BIQ, GRE scores, undergraduate GPA, and first-year graduate school GPA data had to be on file. Of those whose data had been used by the Validity Study Service in the past, only 37 percent had complete data. This resulted in a severe loss of cases, and many studies could not be used. A lower bound of 25 usable cases was required for the inclusion of a particular study. Eighty studies qualified on this ground, but one was subsequently dropped because the standard deviation of self-reported grades was very much smaller for the applicant pool than for the VSS group for that institution. (This is an extremely atypical situation, and one which, when used in the range restriction computations led to a negative estimate of test score variance. Clearly there was something wrong with those data.) Thus, 79 studies were included in the present research. The total number of cases was 3,832, and the study sizes ranged from 25 to 194, with a mean of 48.5 and an interquartile range of 28 to 54.

The history file was searched for the records of all examinees who had had scores sent to the departments whose studies were included in this research. Those for a particular department are referred to elsewhere in this report as the applicant pool for the department.

Analyses

Test analyses for forms 3DGR1, 2, and 3 (Wallmark, 1982a, 1982b, 1982c), which were administered in 1981, provided the data to estimate statistics for all GRE takers. These analyses contained sample sizes and GRE General Test score means, variances, correlations, and reliabilities. The within-administration statistics were used to estimate statistics for the group of all candidates tested in 1981. Since the reliabilities were available, the variance of the error of measurement could be computed for an administration as the test variance times (1 - reliability). A weighted average of these figures was used as the variance of the error of measurement in the total test-taking population.

With the exception of data on the undergraduate school performance of applicants, data on all preadmission variables were available for each VSS group and for each applicant pool. Thus, data were available for three selectors (GRE verbal, quantitative, and analytical scores) in both groups; for one selector (undergraduate school performance) in only the restricted (VSS) group; and for a variable subject to selection (self-reported undergraduate school performance) in both groups. Although this is not the usual configuration of data available for range restriction computations, it was sufficient for estimating the variance of the undergraduate school performance variable and its covariances with scores for the applicant pools (see Appendix B for formulas). Application of the formulas in Appendix B provides estimates of the variance-covariance matrices of selectors--that is, GRE scores and undergraduate school performance--for both the applicant and VSS groups. Data on graduate grade point average was available only for the VSS group. This configuration of data availability, typical in projects that involve correcting for the effects of selection, allowed the use of standard formulas to estimate the applicant pool statistics for the graduate grade point average

(Gulliksen, 1950, pp.165-166; Thorndike, 1982, pp.260-261).

An alternative and much simpler procedure than the one that uses the formulas of Appendix B was considered--that of using the self-reported undergraduate school performances as if they were actual grades. If these variables were highly correlated, this procedures would have been used. However, the relationship between self-reports and actual performances was not high enough. The average correlation of the self-reported with actual undergraduate school performance was only .23, its standard deviation was .15, and its maximum was only .61.

The next step was to estimate test score validities for all GRE General Test takers. In this computation, the applicant pools were regarded as the restricted groups that were selected from all GRE takers. For each department, the validities for all GRE takers were computed. GRE true scores took the role of selectors, with GRE scores and graduate grade point averages being subject to the effects of selection. (As mentioned previously, the motivation for giving the true score this role is put forth in Appendix A.) Because the test statistics for all GRE takers were already computed, it was necessary only to correct the graduate grade point statistics. For this purpose the configuration of information was the same as when generalizing validities to the applicant pool--explicit selector data were available in both the restricted and unrestricted pools, but data on the variable subject to selection were present only in the restricted groups. The correction formulas were the same, but different variables played the roles. Then, because the covariances with true scores were, according to test theory, the same as the covariances with actual test scores, and because the test score statistics for all GRE takers were known, validities for that group could be computed. After this step, covariances and correlations were available for all groups at all levels of selection.

Each hypothesis was evaluated by comparing implied VSS group validities with observed validities. The implied validities were obtained by computing a simplified set of validities, such as the average validities, for the groups in which the generalization was made, and correcting for the effects of selection to obtain the VSS pool statistics for the particular generalization. The generalization hypotheses were each applied at all three levels of selection--VSS group, applicant pool, and all GRE takers.

For the VSS groups, the hypothesis of equal validities was implemented by using the average validity for a measure as its implied validity for each department. The equal ratio hypothesis was tested with the formula given in Appendix C, which multiplies the average validities by a different constant for each department and uses the result as a different set of implied validities.

For the applicant pools, the theoretical validities were found by averaging validities across the pools. Then, using the formula of Appendix D, "reverse" corrections were made for the effects of selection on the test scores and the undergraduate school performance variable to obtain validities implied by the equal validity hypothesis. Another set of implied validities was obtained for the equal ratio hypothesis by applying the ratio-preserving procedures of the formula of Appendix C to the applicant pool validities and again using the reverse correction of Appendix D for the effects of selection on the GRE General Test and undergraduate record.

For generalization at the level of all GRE takers, the validities estimated for all GRE takers were averaged across groups and the averages taken as theoretical validities. Then, using the formula of Appendix D, the reverse correction was applied in two steps to these theoretical validities. The first step accounted for selection on true test scores; the second step accounted for selection on the GRE General Test scores and the undergraduate school performance variable. The two corrections produced implied validities that should be observed in the VSS groups if the generalization hypothesis were true. Another set of implied validities was obtained by applying the ratio-preserving procedures of Appendix C to the validities for the pool of all GRE takers and again applying the two-step correction process.

The results of these computations were evaluated in several ways. First, for each test-hypothesis-group combination, the means and standard deviations for the implied and observed validities were compared, and the correlations between implied and observed validities were computed. Second, the percent of variance of the observed validities that was accounted for by the implied validities and sampling error was calculated. The percent accounted for by the implied validities was simply the square of the correlation between implied and observed validities multiplied by one hundred. The percent of variation of the observed validities accounted for by sampling error was calculated by averaging the error variances of the individual coefficients, dividing the average by the test variance, and multiplying by one hundred. The sample variances of the observed validities were calculated using the same formula used by Pearlman et al. (1980). Third, for the equal ratio hypothesis applied in the pool of all GRE takers, the differences between the implied and observed validities were tested for significance, and the patterns of significance were compared to the types of departments. Finally, the means, standard deviations, fifth percentile values of validities, and percent positive validities were found for the VSS groups, the applicant pools, and the GRE takers. For the applicant pools and the GRE takers, the statistics were obtained for both test scores and true score validities. The fifth percentile was used because, in many validity generalization studies, a figure called the 95 percent credibility value is reported. It is the value above which 95 percent of true score validities are expected over a series of studies.

Results

The statistics for all GRE takers were, for the verbal, quantitative, and analytical scores, respectively, as follows: means were 494, 532, and 516; standard deviations were 123, 133, and 129; standard errors of measurement were 35, 38, and 37; and reliabilities were all .92. The correlations were as follows: .5 between verbal and quantitative, .67 between verbal and analytical, and .7 between quantitative and analytical.

Table 1 contains the means of the validities observed using VSS group data, as well as those estimated under the six generalization hypotheses. Clearly, the generalization hypotheses all led to implied validities that were, on the average, at the right level.

Table 2 contains the standard deviations of the validities observed using VSS group data, as well as those estimated under the six generalization hypotheses. Note that, applied in the VSS groups, the hypothesis of equal validities led to standard deviations that were depressed, but also to better approximations of the observed VSS standard deviations when applied in the applicant and GRE taker groups. The standard deviations of the generalized validities based on the equal ratio hypothesis were not depressed for the VSS groups, but fit well regardless of the groups over which the generalization was made.

Table 3 contains the correlations of the validities observed using VSS group data with those estimated under each generalization hypothesis for each group. Note that the equal validity hypothesis yields implied validities for the quantitative score that have almost no correlation with the observed values, and that the corresponding correlations for the other measures were quite low. The equal ratio hypothesis yields implied validities that correlatc much higher with the observed values, but the correlation declines as successive corrections for range restriction were made.

Another figure of merit for evaluating the success of validity generalization was the percent of variance of observed validity coefficients accounted for by the implied validities and sampling variance. The percents obtained for the equal validity hypothesis are presented in Table 4. Because the average was used in this computation, the VSS group figures of 58, 51, and 64 for the verbal, quantitative, and analytical scores, respectively were the percents of variance of observed validities accounted for by error alone. It can be seen that these quantities are substantial, due probably to the small sample sizes involved. The application of the equal validity hypothesis in the applicant and GRE taker pools afforded little or no improvement over the percent of variance accounted for by sampling error, as one would expect from the very modest correlations for that hypothesis in Table 3. The results for the equal ratio hypothesis were not presented in Table 4 because seven out of nine of them were in excess of one hundred, with the other two also extremely large. These are clearly unacceptable results, which no doubt were obtained because the validities implied by the equal ratio hypothesis were markedly affected by sampling errors; the large correlations for equal ratio hypothesis in Table 3 almost certainly were substantially subject to similar error. The results for the equal validity hypothesis were much less affected by such errors because the amount of overdetermination in calculating the theoretical validities was much less; only three parameters were estimated using the equal validity hypothesis, but 82 (three measures plus 79 institutions) parameters were found for the equal ratio hypothesis.

The observed VSS group validities were tested individually for significant differences from the generalized validities based on the ratio model applied to validities for all GRE takers. With 79 coefficients, one would expect almost four of these tests to reject the null hypothesis at the five percent level. Five of them were significant for the verbal score, four of them were significant for the quantitative score, and three were significant for the analytical score. There were eight patterns of significance and non-significance for the three measures, and six of them were observed over the eight departments where significance was noted. Also, different patterns were noted for departments of the same kind. These results were essentially at the chance level, and with no consistent pattern discernible.

Table 5 contains the means, standard deviations, fifth percentile values of test and true score validities, and percent positive validities for the VSS, applicant, and GRE taker groups. In it can be seen an expected increase in validity as one scans from the restricted VSS groups, through the applicants, to all GRE takers. Note also that true score validities were greater than test score validities, but not greatly so. True score validities were not presented for the VSS groups because the test score reliabilities in those groups were not known. The table shows little difference in the standard deviations of validities. Negative validities exist at all levels of restriction, but by far the greatest majority of coefficients were positive.

Discussion

The context in which validity generalization research arose was that of industrial hiring. Substantial degrees of validity generalization have been reported in this context; i.e., differences in observed validities have been accounted for by statistical artifacts such as restriction of the tests due to their use in hiring, and variation in criterion reliability. Occupations over which generalization has been made cover a wide variety of settings, perhaps even wider than might be encountered across academic institutions, over which one might therefore also expect validity to generalize. This surmise was supported by Linn et al.(1981), who found 70 percent generalization in a study of law school validities. An expectation of the present study was that an even higher percent of variance might be explained if a more complete modeling of the selection procedure were possible using range restriction techniques. Therefore, multivariate corrections for restriction on GRE scores, GRE true scores, and undergraduate school performance were employed. The data were more complete than has usually been the case in such studies, because data on the actual applicant pools were available. In addition, we were able to construct a standardized national population to control the variation in test reliability among groups of applicants. Even so, the generalization hypothesis of equal validities gave a very poor accounting of the observed validities for both the applicant pool and the national pool of GRE takers. A large portion--58 percent, 51 percent, and 64 percent for the verbal, quantitative and analytical measures, respectively--of the variation of observed validity coefficients was due to sampling variation arising from the small sample sizes. In comparison, Boldt (1985) found that 26 percent and 29 percent of variation of coefficients of validity of SAT-V and SAT-M, respectively, was accounted for by sampling, and Linn and Hastings (1984) report 11 percent accounted for in their LSAT study.

In addition to the hypothesis of equal validities, a hypothesis of equal validity ratios was tested for the VSS groups, the applicant groups, and the GRE takers. No association of patterns of significance with discipline was found. This result was not expected, because a difference might reasonably be expected between quantitative and non quantitative disciplines, for example. But Braun and Jones (1985), in an empirical Bayes study of the structure of coefficients of regression of graduate GPA on the GRE General Test, also failed to find differences associated with the discipline. The failure to find a systematic pattern in validities in the present study may have occurred because there is none, but it could also result from the large influence of sampling errors on the implied validities obtained using the equal ratio hypothesis.

The validity generalization work in industry established that site differences in validity were influenced by variations in selectivity and in criterion reliabilities. A major conclusion was that low or negative validities were the exception, implying that a study producing such validities was as suspect as the test involved. Therefore, if a study finds very low or negative validity at a site, this suggests that additional research is needed. Perhaps the criterion needs improving, or perhaps there was a computational or clerical error. The test itself should not be immediately suspect. This industrial research emphasizes a principle that should be more generally appreciated: validity coefficients based on selected incumbents can be poor estimates of a test's actual validity.

In the present study, the validity of the GRE General Test appears to be highly specific, or at least greatly affected by sampling variation. Even so, the thrust of the results of this study coincides with that of the industrial work. The results that provide this thrust are that, even though the lower fifth percentile of the validities ranged from -.01 to -.12 across the three scores, the percent of positive validities was in the high .80s to mid .90s; that is, the great preponderance of validities were positive. The average validities were in the mid-twenties for the VSS groups, rising to the mid-thirties for true scores for all GRE takers. In view of the scarcity of very low or negative validities, studies of the GRE General Test that yield such validities should be questioned.

REFERENCES

American Educational Research Association, Americn Psychological Association, National Council on Measurement in Education (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

American Psychological Association, Division of Industrial-Organizational Psychology (1975). Principles for the validation and use of personnel selection procedures. Dayton, OH: The Industrial-Organizational Psychologist.

American Psychological Association, Division of Industrial-Organizational Psychology (1980). Principles for the validation and use of personnel selection procedures. (2nd edition) Broccoli, CA: Author.

Baird, L. L. (1983). Predicting predictability: The influence of student and institutional characteristics on the prediction of grades. College Board Report No. 83-5. New York: College Entrance Examination Board.

Boldt, R. F. (1985) Generalization of SAT validity across colleges. College Board Report No. 86-3. New York: College Entrance Examination Board.

Braun, H. I. & Jones, D. H. (1985). Use of empirical Bayes methods in the study of the validity of academic predictors of graduate school performance. GRE Board Professional Report GREB No. 79-13P. Princeton, NJ: Educational Testing Service.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice (1978). Adoption by four agencies of Uniform Guidelines on Employee Selection Procedures, Federal Register, 43, 38290-12008.

Graduate Record Examinations Board (1985). Manual for Participation in the Graduate Record Examinations Validity Study Service. Princeton, NJ: Educational Testing Service.

Guion, R. M. (1976). Recruiting, selection and job placement. In M. D. Dunnette (ed.), Handbook of industrial and organizational psychology (pp. 562-575). Chicago: Rand McNally.

Gulliksen, H. (1950). Theory of mental tests. New York: Wiley.

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, 245-259.

Linn, R. L., Harnish, 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.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

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.

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

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

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

Thorndike, R. L. (1982). Applied psychometrics. Boston: Houghton Mifflin.

Wallmark, M. (1982a). Test analysis, aptitude test form 3DGR2. ETS Statistical Report 82-76. Princeton, NJ: Educational Testing Service.

Wallmark, M. (1982b). Test analysis, aptitude test form 3DGR3. ETS Statistical Report 82-64. Princeton, NJ: Educational Testing Service.

Wallmark, M. (1982c). Test analysis, aptitude test form 3DGR1. ETS Statistical Report 82-35. Princeton, NJ: Educational Testing Service.

TABLE 1

MEAN VALIDITIES FOR THE VSS GROUPS, OBSERVED AND IMPLIED, BASED ON THE TWO GENERALIZATION HYPOTHESES APPLIED IN THE THREE SETS OF GROUPS

	VERBAL	QUANT	ANALYT
	OBSERVED IN VSS DATA		
	.23	.24	.28
	IMPLIED VALIDITIES OBTAINED USING HYPOTHESIS AND GROUP INDICATED		
VSS GROUPS			
EQUAL VALIDITIES	.23	.24	.28
EQUAL RATIOS	.23	.24	.28
APPLICANT POOLS			
EQUAL VALIDITIES	.22	.23	.26
EQUAL RATIOS	.22	.24	.26
ALL GRE TAKERS			
EQUAL VALIDITIES	.22	.21	.27
EQUAL RATIOS	.23	.23	.28

TABLE 2

STANDARD DEVIATIONS OF VALIDITIES FOR THE VSS GROUPS, OBSERVED AND IMPLIED, BASED ON THE TWO GENERALIZATION HYPOTHESES APPLIED IN THE THREE SETS OF GROUPS

	VERBAL	QUANT	ANALYT
	OBSERVED IN VSS DATA		
	.19	.20	.18
IMPLIED VALIDITIES OBTAINED USING HYPOTHESIS AND GROUP INDICATED			
VSS GROUPS			
EQUAL VALIDITIES	.00	.00	.00
EQUAL RATIOS	.15	.16	.18
APPLICANT POOLS			
EQUAL VALIDITIES	.25	.15	.16
EQUAL RATIOS	.16	.19	.18
ALL GRE TAKERS			
EQUAL VALIDITIES	.22	.15	.15
EQUAL RATIOS	.17	.19	.19

TABLE 3

VSS GROUP CORRELATIONS BETWEEN THE OBSERVED AND IMPLIED VALIDITIES, BASED ON THE TWO GENERALIZATION HYPOTHESES APPLIED IN THE THREE SETS OF GROUPS

HYPOTHESIS	VERBAL	QUANT	ANALYT
VSS GROUPS			
EQUAL VALIDITIES	.00	.00	.00
EQUAL RATIOS	.83	.78	.90
APPLICANT POOLS			
EQUAL VALIDITIES	.15	.00	.17
EQUAL RATIOS	.67	.65	.82
ALL GRE TAKERS			
EQUAL VALIDITIES	.15	.08	.23
EQUAL RATIOS	.56	.72	.74

TABLE 4

PERCENT OF VSS GROUP VALIDITIES ACCOUNTED FOR BY SAMPLING ERROR AND THE EQUAL VALIDITY HYPOTHESIS APPLIED IN THE THREE SETS OF GROUPS

GROUP	VERBAL	QUANT	ANALYT
VSS GROUPS	58	51	64
APPLICANT POOLS	60	51	67
ALL GRE TAKERS	60	52	69

TABLE 5

MEANS, STANDARD DEVIATIONS, FIFTH PERCENTILE VALUES OF TEST AND TRUE SCORE VALIDITIES, AND PERCENT POSITIVE VALIDITIES, FOR THE VSS, APPLICANT, AND GRE TAKER GROUPS

	VSS GROUPS	APPLICANT GROUPS		GRE TAKERS	
	TEST SCORE	TEST SCORE	TRUE SCORE	TEST SCORE	TRUE SCORE
MEANS					
VERBAL	.23	.27	.29	.32	.34
QUANT	.24	.30	.32	.36	.38
ANALYT	.28	.31	.33	.39	.41
ST. DEV.					
VERBAL	.19	.21	.22	.21	.22
QUANT	.20	.21	.23	.26	.28
ANALYT	.18	.20	.21	.21	.22
5 %-ILE					
VERBAL	-.11	-.08	-.09	-.08	-.08
QUANT	-.13	-.06	-.06	-.12	-.12
ANALYT	-.01	-.02	-.02	-.06	-.06
PERCENT POSITIVE					
VERBAL	85	90	90	92	92
QUANT	90	92	92	87	87
ANALYT	95	94	94	95	95

AFPENDIX A

USE OF TEST THEORY TO REPRESENT THE EFFECTS OF SELF SELECTION

We make a very unrestrictive assumption that self selection and external forces that steer a person towards a particular department can be represented by a vector of variables, and it will be seen that they do not need to be identified. The variables can be represented by the vector variable X . Suppose that, for all GRE-takers, the joint distribution of these variables with a vector variable T of true scores from the subtests of the GRE General Test is a function J(X,T), and assume that errors of measurement are independent of X and T , with distribution D(E). Then, for all GRE-takers, the joint distribution of all these variables is J(X,T)D(E), and the joint distribution of T and E would just be the marginal distribution of T , times D(E).

Now suppose self-selection takes place. By hypothesis, and not a very restrictive one, it occurs by operation of explicit selection on X, and could be represented as G(X)J(X,T), where G adjusts the frequencies according to however the selection worked. Note that selection does not operate explicitly on T , since T cannot be observed. There would be a different G for each institution, and the marginal distribution of T for each institution would be the integral over the space of X of the product of G and J . Since the errors of measurement are independent by hypothesis, the distribution of E would be unaffected. But there would be an adjustment in the distribution of T . Hence the test score distributions would differ only by the distribution of T , and the distribution of E , conditional on T is unaffected and the range restriction formulas apply. Thus X operates on T so that even though T is not an explicit selector, it can take that role in the range restriction formulas because the conditional distributions of E are not affected. In particular, the standard error of measurement is unaffected by the selection and, because the expectation of errors of measurement given true score is zero, the covariance of test scores with true scores and the variance of true scores are equal in both the selected and unselected groups, hence the regression constants are the same in both groups.

Note, as was mentioned above, the really helpful fact that the variables in X need not be known, nor do the forms of J and G.

APPENDIX B

USE OF A SUPPLEMENTARY VARIABLE WHEN DATA ARE MISSING FOR AN EXPLICIT SELECTOR

When capturing data for the routine operations of a secure testing program, for the majority of examinees it is only necessary to obtain test scores and application information. Some examinees, however, may attend institutions that will supply data to the program operator for use in a validity study in which the relationships of test scores, sending institution grades, and receiving institution grades of applicants are studied. If it is desired to estimate validity in an applicant pool, one needs statistics for the explicit selectors in both the applicant and incumbent pools in order to make the needed corrections for the effects of selection. One lacks, however, the sending institution statistics in the applicant pool, and they must therefore be estimated. This can be done if a supplementary variable exists that is present in both the applicant and incumbent pools and that is correlated with the missing explicit selector. This supplementary variable takes the role of a variable subject to the effects of selection. Because it is observed in both the applicant and incumbent pools, it can be used to estimate the missing statistics.

In the present case, the sending institutions are undergraduate schools, the receiving institutions are graduate departments, the incumbents are the graduate students whose data are used, and the applicant pool consists of those who apply to the receiving institutions. The explicit selectors that act on the applicant pool to create the incumbent pool are those of the GRE General Test and undergraduate school grades. The supplementary variable can be a self-reported analog to the undergraduate school grade since a test program can easily collect examinee-reported biographical information on the registration form.

The range restriction assumptions are that the coefficients of regression of variables subject to selection on the explicit selectors are undisturbed by the selection process, as are the errors of prediction of the variables subject to selection by the explicit selectors. Therefore, the following normal equations for estimating regression coefficients in the applicant pool are satisfied by regression coefficients calculated in the incumbent pool.

$$C_{vs} = C_{vv} B_v + C_{vq} B_q + C_{va} B_a + C_{vp} B_p \qquad (1)$$

$$C_{qs} = C_{vq} B_v + C_{qq} B_q + C_{qa} B_a + C_{qp} B_p \qquad (2)$$

$$C_{as} = C_{av} B_v + C_{aq} B_q + C_{aa} B_a + C_{ap} B_p \qquad (3)$$

The symbols v, q, a, p, and s stand for verbal, quantitative, analytic, actual undergraduate school grade, and self-reported undergraduate school grade, respectively. C_{uv} is the covariance of u and v calculated in the applicant pool and B_x is the coefficient of partial regression of s on x calculated in the incumbent pool, hence known. Further, all covariances for which both variables are observed by the test program in the applicant pool are known. This leaves C_{vp}, C_{qp} , and C_{ap} as the only unknown quantities in equations (1), (2), and (3) respectively. To find them use

$$C_{vp} = (C_{vs} - C_{vv} B_v - C_{vq} B_q - C_{va} B_a)/B_p \qquad (4)$$

$$C_{qp} = (C_{qs} - C_{vq} B_v - C_{qq} B_q - C_{qa} B_a)/B_p, \text{ and} \qquad (5)$$

$$C_{ap} = (C_{as} - C_{va} B_v - C_{qa} B_q - C_{aa} B_a)/B_p \qquad (6)$$

From the assumption that the errors of prediction are unaffected by the selection process, we obtain

$$C_{ss}=c_{ss}+b'(||C_{xx}||-||c_{xx}||)b \qquad (7)$$

where ||Cxx|| and ||cxx|| are the explicit selector variance-covariance matrices in the applicant and incumbent pools, respectively, and b is a column vector of partial regression coefficients, the B_x . With the computations of equations (4) through (6) completed, all the quantities for equation (7) are available except the variance of the sending institution grade in the applicant pool, C_{pp} . If we define $||C_{xx}^*||$ as being the same as $||C_{xx}||$ but with a zero in the position of C_{pp} , which is third column and third row, then (5) becomes

$$C_{ss}=c_{ss}+b'(||C_{xx}^*||-||c_{xx}||)b+C_{pp}(B_p^2) \qquad (8)$$

All quantities in (8) are known except C_{pp} , for which the solution is

$$C_{pp}=(C_{ss}-c_{ss}-b'(||C_{xx}^*||-||c_{xx}||)b)/(B_p^2) \qquad (9)$$

With the calculation of C_{pp} in equation (9), all the entries in $||C_{xx}||$ are available.

APPENDIX C

GENERALIZING THE ASSUMPTION THAT THE VALIDITIES ARE PROPORTIONAL ACROSS INSTITUTIONS

According to the hypothesis that the ratios of the validities of the subtests of the GRE General Test are the same across institutions, the validity, V_{it} , of a subtest, t , for institution i , is the product of F_t , a constant associated with the subtest, and G_i , a constant associated with the institution. Then

$$V_{it} = F_t G_i + \text{error} \qquad (1)$$

Neglecting the error,

$$V_{\cdot t} = F_t G_{\cdot} \qquad (2)$$

where the dot indicates averaging over the missing variable, i in this case. Then, using (1) and (2)

$$V_{it}/V_{\cdot t} = G_i/G_{\cdot} \qquad (3)$$

for any of the subtests. Therefore,

$$(V_{iv}/V_{\cdot v}+V_{iq}/V_{\cdot q}+V_{ia}/V_{\cdot a})/3=G_i/G_{\cdot}=K_i \qquad (4)$$

Then, from equation (2),

$$V_{\cdot t} K_i = F_t G_i \qquad (5)$$

Thus, equation (5) gives the formula for estimating the validity of the test under the hypothesis that the ratios of the validities of the subtests are the same for all institutions. Since the average test validities are both multiplied by the same value, K_i , their ratio is the same for all groups, and their level varies with variation in the magnitude of K_i .

APPENDIX D

CALCULATING VALIDITIES IN THE RESTRICTED GROUP

In the present study, once the generalization has been made, we want to reverse the range restriction calculations from, the population of all GRE takers to an applicant pool, or from an applicant pool to a VSS group. In this situation, the validity in the unselected group is known, but not the validity in the restricted group, one would wish to have a restricted criterion variance that is consistent with the restricted validity. If we regard as our second set of unknowns the ratio of the criterion variances, there is enough information to solve the problem. This can be seen as follows. The assumption that selection does not affect the regression function can be written as follows:

$$||C_{xx}||^{-1}||C_{xy}|| = ||c_{xx}||^{-1}||c_{xy}|| = b \qquad (1)$$

where $||C_{uv}||$ and $||c_{uv}||$ are covariance matrices for variables u and v for the unrestricted and restricted populations, respectively. The x variables are the explicit selectors, y is subject to selection and is a single variable in this project. Therefore the covariances involving both x and y are arranged in a column vector, as is b. Equation (1) can be rewritten as

$$||M||S_y = ||c_{xx}||^{-1}||s_x|| \; ||r_{xy}||s_y \; , \qquad (2)$$

where $||M|| = ||C_{xx}||^{-1}||S_x|| \; ||R_{xy}||$,

S_y and s_y are the standard deviations of y for the unrestricted and restricted populations, respectively; $||S_x||$ and $||s_x||$ are diagonal matrices of standard deviations of the explicit selectors for the restricted and unrestricted populations, respectively; and $||R_{xy}||$ and $||r_{xy}||$ are the correlations between explicit selectors and y for the unselected and selected populations, arranged as column vectors. The assumption that selection does not affect the errors of prediction of y by x can be written as follows:

$$S_y^2 = s_y^2 + b'(||C_{xx}||-||c_{xx}||)b \; ,$$

and hence as

$$S_y^2 = s_y^2 + ||M||'(||C_{xx}||-||c_{xx}||)||M|| \; S_y^2 \; .$$

Therefore,

$$(S_y/s_y) = 1/(1-||M||'(||C_{xx}||-||c_{xx}||)M)^{.5} = F\ . \qquad (3)$$

Then using (2), and the definitions of $||M||$ and F ,

$$||s_x||^{-1}||c_{xx}||\ ||M||\ F = ||r_{xy}||\ . \qquad (4)$$

All the information needed to carry out the calculation indicated on the left hand side of equation (4) is known after the generalized unrestricted validities are known. The development of this computation used the range restriction assumptions to arrive at restricted correlations without using standard deviations of y , in order to obtain estimates consistent with the assumptions.