Psychological Reports, 1993, 72, 123-127.

VALIDITY OF THE GRE WITHOUT RESTRICTION OF RANGE[1]

BRADLEY E. HUITEMA AND CHERI R. STEIN

Western Michigan University

Summary.—Restriction of range is a frequently acknowledged issue in estimating the validity of predictors of academic performance in graduate school. Data obtained from a doctoral program in a psychology department where graduate students were admitted without regard to Graduate Record Examination (GRE) scores yielded essentially identical standard deviations on this test for the 204 applicants and 138 enrolled students. The GRE-Total validity coefficients obtained on subjects in the enrolled sample ranged from .55 through .70; these values are considerably higher than those typically reported. The data are congruent with the argument that uncorrected GRE validity coefficients yield biased estimates of the unknown validity in unrestricted applicant pools.

The Graduate Record Examinations General Test (Educational Testing Service, 1949-1986) is perhaps the most widely used, validated predictor of academic performance at the graduate level. Most studies report fairly low validity coefficients for the Verbal and Quantitative sections (both individually and combined) of the GRE, regardless of academic department and method of measuring academic achievement. Values ranging from about .20 through the low .30s are typical for studies in which first-year graduate GPA is the criterion variable. There is, however, much variation across studies. Earlier validation studies summarized by the Educational Testing Service (1977) showed median validity coefficients for nine major fields ranging from .02 to .36 for the Verbal section and from .06 to .32 for the Quantitative section. Many individual studies have yielded values somewhat beyond these ranges (e.g., Thornell & McCoy, 1985).

Differential curricula, measurement error (especially in the criterion variable), function used in fitting the data, sampling bias, sampling error, and various forms of scaling error are among the many issues that should be considered when attempting to explain why validity coefficients tend to be low and why they vary across studies. While some of these issues are relevant to only subsets of the entire body of published work in this area, an easily recognized form of sampling bias, widely known as restriction of range, appears to afflict virtually all published validation studies of the GRE.

Corrections commonly used for problems of restriction in range were developed almost a century ago and have been discussed in the classic psychometric textbooks (e.g., Gulliksen, 1950; Lord & Novick, 1968; Thorndike, 1949) for many decades. More recent work (e.g., Braun & Szatrowski,

[1]Address correspondence regarding this article to Bradley E. Huitema, Department of Psychology, Western Michigan University, Kalamazoo, MI 49008. Cheri R. Stein is now at Training Concepts, South Haven, MI.

1984a, 1984b; Gross & Fleischman, 1983; Olson & Becker, 1983) has yielded procedures that are appropriate for a wide variety of applications under certain assumptions, but most researchers simply report the uncorrected (but biased) validity coefficients and point out that they are probably underestimates of the coefficients that would have been obtained had the unrestricted population been sampled. When such coefficients are relatively close to zero, as is frequently the case in the literature on GRE validation, it is often difficult to argue that restriction of range is the major reason for obtaining a low validity coefficient because other issues such as sampling error may appear to be equally persuasive.

If data from samples of students who were not selected on the basis of the test were available, more convincing estimates of the validity of the GRE in predicting graduate school achievement in the unrestricted applicant pool could be obtained. Validity coefficients based on data of this type would not require the support of correction formulas, all of which rest on strong assumptions (Brown, Stout, Dalessio, & Crosby, 1988). Generally such data have not been available because academic departments in which GRE scores are required for admission have not been willing to adopt a policy to ignore the test data in the admissions process. This was not the case, however, in the present study.

This curious state of affairs came about because the unwritten but *de facto* graduate admissions policy that operated in the Department of Psychology at Western Michigan University from the early 1970s through the mid-1980s was to ignore GRE scores even though they were required by the Graduate College for admission to doctoral programs. This practice evolved from the dual views held by some faculty that (1) the use of such tests was elitist and had no place in a public institution and (2) the teaching technology used by most faculty in the department was adequate to produce mastery of course content irrespective of individual differences in entry skill. A benefit of this admissions policy was that it made possible the collection of data under a nearly ideal validation paradigm.

Method

The doctoral programs in applied and experimental areas received 204 completed applications from the 1970s (when the programs were initiated) through 1985-86; 138 of these applicants enrolled. Predictor data collected on all applicants included undergraduate GPA and GRE Verbal and Quantitative scores. Academic performance data were collected on all enrolled students ($n = 138$); these data were in the form of points earned on multiple examinations administered in three graduate courses (Advanced Statistics, Assessment Methods, and Research Methods in Applied Behavior Analysis) and an Over-all Evaluation score assigned by a committee of four graduate faculty. The steps involved in obtaining the Over-all Evaluation scores in-

cluded (1) providing the evaluation committee with a list of the names of the 138 current and former graduate students, (2) providing a description of the "ideal Psychology doctoral student" (developed from departmental mission statements), and (3) asking each of the four faculty to identify (a) the 25 students who most closely matched the ideal and (b) the 25 students who least matched the ideal.

Fifty students were identified by one or more of the committee members as closely matching the ideal and 44 were identified by at least one faculty member as least like the ideal; another 44 were not identified for either of the extreme categories. Scale values were attached to the three categories (ideal = 3, neither = 2, and least ideal = 1) to provide an Over-all Evaluation score for each student.

Validity coefficients based on the best fitting regression models were computed using undergraduate GPA and GRE-Total scores as predictors. Points earned in each of the three graduate courses and the Over-all Evaluation scores served as the criterion measurements.

Graduate grade point average was not employed as a measure even though it has been the most frequently employed criterion in GRE validation studies. It was not used because the grading approach was criterion-referenced rather than norm-referenced. The grade "A" was assigned in most courses if a specified proportion of the course content was "mastered." Repeated "remedial" testing was allowed in most courses if initial examination scores were below the "A" level. This educational approach yielded a severe ceiling effect; there was virtually no grade variation. Interestingly, all faculty agreed that there was much variation in the academic skill of the students; the assigned grades simply did not reflect this variation. The ceiling effect did not occur with the criteria of points earned on initial (rather than remedial) course examinations and the Over-all Evaluation score. Indeed, the motivation for using measures other than the graduate GPA was to avoid the problem of restriction of range on the criterion variable.

Results

Table 1A provides evidence that restriction of range was not characteristic of GRE data in the selected sample. Notice that the GRE standard deviations for all applicants differ very little from the standard deviations associated with those who enrolled. In the case of Total scores (i.e., combined Verbal and Quantitative scores) the standard deviations are essentially identical.

Table 1B contains the GRE-Total validity coefficients associated with each of the four measures of academic performance. It can be seen that the four coefficients are fairly homogeneous, statistically significant, and quite high relative to those found in other GRE validation studies. Also, note that the best-fitting model was not always a simple linear function. A quadratic

TABLE 1
GRE: STANDARD DEVIATIONS AND VALIDITY COEFFICIENTS

A. Standard Deviations for All Applicants and Enrolled Doctoral Students

GRE Scale	All Applicants $n = 204$	Enrolled Doctoral Students $n = 138$
Verbal	105.36	109.78
Quantitative	108.71	105.08
Total	188.89	188.27

B. GRE Total Validity

Criterion	Type of Coefficient	Coef.	n	F	df	p
Advanced Statistics Course Points	Correlation Index R_I (based on quadatic model)	.60	107	29.25	2,104	<.00001
Behavioral Assessment Course Points	Pearson r	.70	30	26.90	1,28	<.00002
Research Methods in Applied Behavior Analysis Course Points	Pearson r	.55	42	17.35	1,40	<.0002
Over-all Evaluation Classification	Triserial r	.63	138	34.53	1,136	<.00001

model was required to fit the GRE scores adequately to the Advanced Statistics course data. Achievement in this course improved more rapidly as GRE scores increased from very low to intermediate values than when they increased from intermediate to high values.

Correlations between undergraduate GPA and the four criteria (not reported here) were all nonsignificant. Further, regression analyses including both undergraduate GPA and GRE Total scores in combination yielded no advantage in retaining GPA in the model. It is concluded that for the population sampled, undergraduate GPA is not a valid predictor individually and that it does not provide incremental validity. It should be pointed out, however, that undergraduate grade inflation for WMU students in the sample is the likely explanation for this finding of very low undergraduate GPA validity.

DISCUSSION

The validity coefficients reported here are considerably higher than those typically encountered in the literature on GRE validation, regardless of criteria employed. This finding is consistent with the argument that range restriction produces validity coefficients that underestimate GRE validity in the unrestricted applicant population. Since restriction of range was not an issue in this study, higher than typical values should be expected.

Additional evidence of the potential effect of restriction of range can be demonstrated by truncating the GRE distribution at some typical cut-off score and computing the validity on the retained data. Suppose, for example,

that all admitted subjects having GRE-Total scores below 1200 (which is a typical cut-off score for selective programs) had not been admitted and were not included in the sample. The GRE-Total validity (using Over-all Evaluation scores as the criterion) is only .24 for the subsample of those with GRE-Total scores of at least 1200, whereas the complete (unrestricted) sample validity is .63. The former coefficient (.24) is within the range of typical values reported in earlier reviews of GRE validity in predicting first-year performance.

In summary, the observed validity coefficients based on the unrestricted sample are consistent with values predicted from correction for restriction of range formulas applied to typical coefficients from previously published studies that employed restricted samples. Likewise, validity based on intentionally restricted observed data is consistent with values predicted by statistical theory. These results support the conventional argument that uncorrected GRE validity estimates based on range-restricted samples are strongly biased toward zero.

REFERENCES

BRAUN, H. I., & SZATROWSKI, T. H. (1984a) The scale-linkage algorithm: construction of a universal criterion scale for families of institutions. *Journal of Educational Statistics*, 9, 311-330.

BRAUN, H. I., & SZATROWSKI, T. H. (1984b) Validity studies based on a universal criterion scale. *Journal of Educational Statistics*, 9, 331-344

BROWN, S. H., STOUT, J. D., DALESSIO, A. Y., & CROSBY, M M. (1988) Stability of validity indices through test score ranges. *Journal of Applied Psychology*, 73, 736-742.

EDUCATIONAL TESTING SERVICE. (1949-1986) *Graduate Record Examinations Aptitude Test.* Princeton, NJ: Author.

EDUCATIONAL TESTING SERVICE. (1977) *Graduate Record Examinations technical manual.* Princeton, NJ: Author.

GROSS, A., & FLEISCHMAN, L. (1983) Restriction of range corrections when both distribution and selection assumptions are violated. *Applied Psychological Measurement*, 7, 227-237.

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

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

OLSON, C. A., & BECKER, B. E. (1983) A proposed technique for the treatment of restriction of range in selection validation. *Psychological Bulletin*, 93, 137-148.

THORNDIKE, R. L. (1949) *Personnel selection.* New York: Wiley.

THORNELL, J. G., & MCCOY, A. (1985) The predictive validity of the Graduate Record Examinations for subgroups of students in different academic disciplines. *Educational and Psychological Measurement*, 45, 415-419.

Accepted November 23, 1992.