Psychological Bulletin
2001, Vol. 127, No. 1, 162–181

0033-2909/01/$5.00 DOI: 10.1037//0033-2909.127.1.162

A Comprehensive Meta-Analysis of the Predictive Validity of the Graduate Record Examinations: Implications for Graduate Student Selection and Performance

Nathan R. Kuncel
University of Minnesota, Twin Cities Campus,
and Personnel Decisions International

Sarah A. Hezlett
University of Minnesota, Twin Cities Campus,
and Questar Data Systems

Deniz S. Ones
University of Minnesota, Twin Cities Campus

This meta-analysis examined the validity of the Graduate Record Examinations (GRE) and undergraduate grade point average (UGPA) as predictors of graduate school performance. The study included samples from multiple disciplines, considered different criterion measures, and corrected for statistical artifacts. Data from 1,753 independent samples were included in the meta-analysis, yielding 6,589 correlations for 8 different criteria and 82,659 graduate students. The results indicated that the GRE and UGPA are generalizably valid predictors of graduate grade point average, 1st-year graduate grade point average, comprehensive examination scores, publication citation counts, and faculty ratings. GRE correlations with degree attainment and research productivity were consistently positive; however, some lower 90% credibility intervals included 0. Subject Tests tended to be better predictors than the Verbal, Quantitative, and Analytical tests.

Effective selection and training of graduate students is of critical importance for all fields requiring graduate training. Admission of poorly qualified students misuses the resources of students, faculty, and schools. Failure to admit and retain outstanding candidates ultimately weakens a field. Standardized tests, especially the Graduate Record Examinations (GRE), have been heavily weighted sources of information in admission decisions for many departments. The GRE, published by the Educational Testing Service (ETS), is a set of standardized tests designed to predict the scholastic performance of graduate students. The GRE includes tests of verbal, quantitative, and analytic abilities as well as tests of subject area knowledge for a number of fields.

In their review of selection data examined by psychology admission committees, Norcross, Hanych, and Terranova (1996) reported that GRE scores are required by 93% of doctoral programs and 81% of master's programs. In addition, the GRE is often used to help decide which students will receive fellowships and other awards. Although the weight given to this instrument for admission decisions varies from university to university, most highly competitive institutions have high minimum score requirements (Norcross et al., 1996). Given their widespread use, the validities of the GRE General and Subject Tests for prediction of graduate school performance are clearly important.

The GRE was specifically designed to measure "basic developed abilities relevant to performance in graduate studies" (Briel, O'Neill, & Scheuneman, 1993, p. 1). The test items reflect long-term learning of material related to graduate performance. On the General Test, test takers are asked to solve problems, synthesize information, and resolve sometimes complex relationships between pieces of information. Specifically, the Verbal measure (GRE-V) contains analogy, antonym, sentence completion, and reading comprehension problems. The Quantitative measure (GRE-Q) is composed of discrete quantitative, quantitative com-

Nathan R. Kuncel, Department of Psychology, University of Minnesota, Twin Cities Campus, and Personnel Decisions International, Minneapolis, Minnesota; Sarah A. Hezlett, Department of Psychology, University of Minnesota, Twin Cities Campus, and Questar Data Systems, Eagan, Minnesota; Deniz S. Ones, Department of Psychology, University of Minnesota, Twin Cities Campus.

An earlier version of this article was presented at a symposium conducted at the 13th Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, Texas, April 1998, and portions of the article were presented at the 106th Annual Convention of the American Psychological Association, August 1998, San Francisco, California.

We thank Tom Bouchard, John P. Campbell, Matt McGue, Frederick L. Oswald, John Ruscio, Frank Schmidt, Kenneth Sher, Auke Tellegen, C. Viswesvaran, and Steffanie Wilk for their helpful comments and suggestions; Barton Adams, Brian Griepentrog, Yoshani Keiski, and Jeanette Shelton for their assistance in gathering the articles summarized in this meta-analysis; and Rob Durso and Tom Jerilie for providing raw Educational Testing Service validity study data. Nathan K. Kuncel gratefully acknowledges the National Science Foundation for indirect support for this project through a Graduate Research Fellowship.

Correspondence concerning this article should be addressed to Nathan R. Kuncel, Department of Psychology, University of Minnesota, Twin Cities Campus, N218 Elliott Hall, 75 East River Road, Minneapolis, Minnesota 55455-0344. Electronic mail may be sent to kunce001@tc.umn.edu.

parison, and data interpretation problems. The Analytical measure (GRE-A) includes analytical reasoning and logical reasoning items. The Subject Tests assess acquired knowledge specific to a field of study (e.g., biology, chemistry, or psychology; Briel et al., 1993).

Numerous studies of the GRE's validity have been conducted, with papers appearing soon after the tests were developed in the 1940s (e.g., Cureton, Cureton, & Bishop, 1949). The results of this half century of research have been inconsistent and controversial. Although some researchers concluded that the GRE General and Subject Tests are valid predictors of graduate school performance (e.g., Broadus & Elmore, 1983; Sleeper, 1961), others found small relationships between GRE scores and success in graduate school (e.g., Marston, 1971; Sternberg & Williams, 1997). Reported validities for the GRE have ranged between −.62 and .81. Given these variable results, doubts about using the GRE to predict graduate school performance have been raised for multiple disciplines, ranging from physics (Glanz, 1996) to journalism (Brown & Weaver, 1979).

Given the volume and nature of the research on the GRE, it is not surprising that several previous summaries and meta-analyses have been conducted (Goldberg & Alliger, 1992; Morrison & Morrison, 1995; Schneider & Briel, 1990). Goldberg and Alliger (1992) meta-analyzed the validities of the GRE for psychology graduate programs, cumulating results across 10 studies. They obtained a correlation of .15 for both the GRE-V and GRE-Q in predicting graduate grade point average (GGPA; N = 963). Morrison and Morrison (1995) obtained similar but slightly larger correlations in their meta-analysis of 22 studies on predicting GGPA in various fields. The GRE-V and GRE-Q displayed correlations of .28 and .22 with this criterion. Consequently, researchers remained critical of the GRE, stating that the observed average correlation was too small to be of use in prediction. Schneider and Briel (1990) performed a quantitative review of validation studies conducted by ETS and reported correlations between GRE scores and 1st-year GGPA of .18 to .32.

Our study improved on previous reviews and meta-analyses in three major ways. First, unlike previous meta-analyses that have focused on either a single population (i.e., psychology graduate students; Goldberg & Alliger, 1992) or criterion measure (i.e., grade point average [GPA]; Morrison & Morrison, 1995), this study examined the validity of the GRE for multiple disciplines using multiple criterion measures. We cumulated results across 1,521 studies. The number of correlations that contributed to our database was 6,589. In contrast, Goldberg and Alliger (1992) meta-analyzed 97 correlations from 10 studies, and Morrison and Morrison (1995) meta-analyzed 87 correlations from 22 studies. Second, all previous reviews and meta-analyses have not directly addressed statistical artifacts that attenuate the magnitude of the relationship between the GRE and graduate school performance measures. This is a major shortcoming, because statistical artifacts such as range restriction and unreliability in criteria attenuate correlations between the GRE and relevant criteria (Kuncel, Campbell, & Ones, 1998). Third, this meta-analysis included an examination of the validity of multiple predictors (e.g., GRE subtest scores and undergraduate GPA [UGPA]) used in combination to predict graduate school performance. Thus, this meta-analysis provides more accurate estimates of the validity of the GRE across disciplines and criteria in addition to new information on the validity of combinations of often-used predictors.

Given the volume of research, the apparently inconsistent results across studies, and strong opinions on both sides about the usefulness of the GRE in predicting graduate student performance, a comprehensive meta-analysis of the GRE's validity is necessary. To thoroughly investigate the validity of the GRE, three aspects of the validation need to be addressed: theoretical, statistical, and methodological. Previous criticisms of GRE validation research have also centered on these three areas. We discuss each of these issues in turn.

Determinants of Graduate School Performance: A Theoretical Argument

Theoretical criticisms of previous GRE validation studies have argued that the GRE does not capture all relevant abilities. Furthermore, these criticisms have pointed to the fact that most validation studies of the GRE have been atheoretical and have not addressed the question of why the GRE should predict graduate school performance.

Past research from the work performance domain suggests that cognitively loaded performance measures can be predicted by measures of general cognitive ability (Hunter & Hunter, 1984). General cognitive ability has demonstrated moderate to large relationships with performance measures in low, medium, and high complexity occupations, respectively (Hunter, 1980). Because the GRE-V, GRE-Q, and GRE-A are similar to many measures of general cognitive ability, scores on these tests should predict performance in an academic setting. Nevertheless, the magnitude of the relationship between the GRE and a dimension of graduate performance depends on how the latter is measured, particularly the extent to which the performance is determined by cognitive ability.

An explanation for the relations between ability and later performance can be found in two sets of research streams: (a) those examining the determinants of work performance (Campbell, Gasser, & Oswald, 1996; McCloy, Campbell, & Cudeck, 1994) and (b) those investigating the relationships among general cognitive ability, job knowledge, and work performance (Borman, Hanson, Oppler, Pulakos, & White, 1993; Schmidt & Hunter, 1993; Schmidt, Hunter, & Outerbridge, 1986). We discuss these two streams of research in turn, focusing on theoretical explanations.

McCloy et al. (1994) demonstrated that performance can be conceptualized as a function of declarative knowledge, procedural knowledge, and motivation. Declarative knowledge is defined as understanding what to do. Procedural knowledge is knowing how to do a task. And motivation is the decision to act, the intensity of the action, and the persistence of action. McCloy et al. empirically demonstrated that individual-differences variables (e.g., general cognitive ability and conscientiousness) affect performance indirectly through their influence on declarative knowledge, procedural knowledge, or motivation. The GRE-V, GRE-Q, and GRE-A quantify individual abilities or skills that would have an influence on later graduate performance through declarative or procedural knowledge. For example, reading and summarizing a passage (GRE-V) would be an example of procedural knowledge relevant for some graduate school performances. Because the GRE is a predictor of maximal performance, theoretically, it will be primarily a determinant of declarative and procedural knowledge and capture few individual differences in motivation.

A second stream of research examined the relations among general cognitive ability, job knowledge, and performance in a number of meta-analytic and large-scale individual studies (Borman et al., 1993; Schmidt & Hunter, 1993; Schmidt et al., 1986). All of the studies arrived at similar conclusions. General cognitive ability was found to have the strongest direct relationship with job knowledge, which suggested that general cognitive ability is related to the acquisition of job knowledge. Job knowledge, in turn, was most strongly associated with job performance, either measured with a maximal performance measure through a work sample or a typical performance measure through supervisor performance ratings. Finally, general cognitive ability had positive and strong relationships with work sample tests. These findings are largely consistent with the work of McCloy et al. (1994) in that general ability has its influence on job performance variables through job knowledge (declarative knowledge) and work sample performance (procedural and declarative knowledge). Further, multiple studies examining relationships between cognitive ability and training success in work settings have also typically revealed large correlations (e.g., Hirsh, Nothrup, & Schmidt, 1986; Hunter, 1980; Pearlman, Schmidt, & Hunter, 1980; Schmidt, Hunter, & Caplan, 1981; Schmidt, Hunter, Pearlman, & Shane, 1979). These studies provide further evidence that cognitive ability predicts acquisition of job knowledge, indicated by success in training.

On the basis of these large bodies of work, one would expect the GRE-V, GRE-Q, and GRE-A to be correlated with graduate performance, especially with criteria that are the academic equivalent of job knowledge. Given that job knowledge has a more direct and stronger relationship with subsequent performance than does general cognitive ability, the GRE Subject Tests are likely to be a better predictor of graduate school performance than the General Tests. All else equal, one would expect that a student entering graduate school with more "job" knowledge would perform better than one who had less "job" knowledge. The student with greater job knowledge would have a better framework into which to integrate field-specific knowledge, enhancing learning. Even without any additional learning, the student with greater job knowledge would perform better in classes and on comprehensive exams, write a better dissertation, perform better as a teaching or research assistant, and generate better research than a student with a lower level of job knowledge.

Role of Statistical Artifacts in GRE Validation Research

Statistical concerns have frequently been raised about previous studies of the GRE's validity. These concerns have revolved around restriction of range, criterion unreliability, and inadequate sample size. This meta-analysis directly addressed each of these statistical issues.

Range restriction and criterion unreliability attenuate the observed correlations between GRE scores and performance in graduate school. Estimates of the GRE's validities are based on individuals who already have been admitted to and attended a graduate program. Given the goal of selecting the most capable students from the applicant pool, the validity coefficients of interest are those based on the applicant group. Because many programs explicitly use the GRE to make admission decisions, it is likely that the range of GRE scores for graduate school incumbents is smaller than the range for graduate school applicants. Restriction of range results in underestimates of GRE validity coefficients for the actual applicant populations. Although many researchers have noted this problem, previous studies typically have not estimated the extent to which range restriction attenuates GRE validity coefficients (a few notable exceptions include House, 1983; Huitema & Stein, 1993; Michael, Jones, & Gibbons, 1960; and Oldfield & Hutchinson, 1997).

Similarly, it has been noted that unreliability in measures of graduate school performance attenuates observed GRE validity coefficients. Measurement error masks the magnitude of the correlation between predictors and criterion constructs. It would be inappropriate in this situation to correct for predictor unreliability because admission committees must base their decisions on GRE scores (unreliability and all). It is appropriate, however, to correct for measurement error in the criteria, because the object is to evaluate how well actual performance, not performance obscured by unreliability, is predicted. The present meta-analysis included corrections for both range restriction and criteria unreliability.

Methodological Issues in Defining Criteria: What Is Graduate School Performance?

The methodological criticisms of previous research have also focused on the inadequacy of most individual criterion measures. This meta-analysis examined eight different criteria: (a) GGPA, (b) 1st-year GGPA, (c) comprehensive examination scores, (d) faculty ratings, (e) number of publications–conference papers, (f) number of times publications are cited, (g) degree attainment, and (h) time to degree attainment. In the following paragraphs, we discuss the relevance and relative importance of these criteria in graduate work. Note that the criteria are largely measures of graduate school performance rather than measures of more distal career success. Although graduate school success is likely to be associated with later career success, the GRE was developed and is used to predict the former and not the latter.

Previous research has suggested that graduate school performance is multidimensional (Enright & Gitomer, 1989; Reilly, 1974). Extending Campbell's model of work performance (Campbell, 1990; Campbell et al., 1996) to the graduate school setting (Campbell, Kuncel, & Oswald, 1998), one would expect these dimensions of performance to be determined by certain sets of declarative knowledge, procedural knowledge, and motivation.

Virtually no previous research on predicting graduate school performance has explicitly distinguished between different dimensions of performance. Yet, the criterion measures used in past research capture different aspects of these broad constructs. Consequently, choices of different criterion measures have resulted in different implicit choices of relevant performance determinants. For example, we would not expect a strong correlation between the GRE and a criterion measure largely determined by motivation (e.g., number of hours studied per week or persisting when it is "cold, wet, or late" [Campbell et al., 1996]). The GRE-V and GRE-Q are primarily tests of ability rather than measures of interests, persistence, or motivation. Another measure (such as an interest inventory or a statement of intent) may be a better predictor of motivationally determined aspects of graduate student performance.

Criteria that are events or outcomes not largely under the control of the graduate student tend to be poorer measures of performance. What aspect of graduate performance the criterion measure captures and how well this aspect is measured will also influence the relationship between GRE scores and graduate school performance. Criteria that are poor measures of any dimension of graduate performance are less likely to be strongly related to scores on the GRE. Also, as in measuring work performance, criterion relevance, accuracy, deficiency, and reliability are important considerations.

GGPA and 1st-year GGPA are the most widely used measures of graduate school performance. GGPA has a number of advantages and disadvantages as a criterion measure. In its favor, GGPA measures long-term work, knowledge acquisition, effort, persistence, and ability. It is also related to post-school success (Hoyt, 1966; Roth, BeVier, Switzer, & Schippmann, 1996). Not favoring GGPA is the fact that grading standards can vary widely across schools, departments, and even faculty teaching the same course (Hartnett & Willingham, 1980). Because of these criterion considerations, we expect moderate correlations between GRE scores and GGPA. As mentioned earlier, because previous research has demonstrated a stronger relationship between job knowledge and subsequent performance than between general cognitive ability and performance, we also expect the GRE Subject Tests to have larger validities than the General Tests with this criterion.

Comprehensive examination scores are also a key part of graduate work. Passing such exams often represents an important stage in graduate school progress and is an important indicator that students are mastering the necessary material. Much like GGPA, several aspects of comprehensive examinations can vary across programs, including examination difficulty, grading standards, and relevance to graduate progress. Most institutions use comprehensive examinations to assess the degree of job knowledge (declarative knowledge) amassed by graduate students. For example, knowledge of psychology or biology is necessary for performance as a psychologist or biologist, respectively. Hence, we expect moderate correlations between GRE scores and comprehensive examination scores, with larger correlations for the Subject Tests.

Faculty ratings can be used to measure a wide range of graduate student characteristics. This study included only ratings that were judged to be related to graduate school performance. If the graduate program included outside-of-school tasks related to graduate school performance (e.g., counseling effectiveness or internship performance ratings), these ratings were also included. If validities for multiple ratings were given, the correlations were averaged for each predictor. Measures that were omitted because of criterion relevance considerations included ratings of friendliness, life balance, and empathy. We also excluded ratings that specifically targeted class performance. This was done to help differentiate faculty ratings from GGPA. Ratings have the advantage of being flexible and can be developed to cover content areas. On the other hand, they can fall prey to the personal biases of the rater as well as rating errors including halo, central tendency, and other response sets (Cronbach, 1990). When large numbers of ratings are obtained for a large number of students, the heavy demand on faculty time becomes severe, resulting in poor discriminant validity. Given that we included only ratings of overall performance, internship performance, and research work in this study, we expect moderate to high correlations between the GRE and faculty ratings.

Research productivity, as indicated by the number of publications or conference papers a student produces either during or after graduate school, has clear links to scientific productivity, which is often a goal of research-oriented programs. However, there are some major disadvantages to this criterion. First, many schools plan on training both scientists and practitioners. Thus, this criterion may simply not apply to a majority of students pursuing pure teaching or applied careers. Second, although quality and quantity are positively related, their intercorrelation is less than perfect (Viswesvaran, 1993). To the extent that the number of publications and conference papers produced represents an interest in research and persistence in the journal review process, we expect relatively low but positive correlations between graduate students' scores on an ability measure such as the GRE and number of publications. Among a group of research-oriented students, the GRE would be more likely to differentiate between those who are more and less successful. Another drawback of using number of publications as a criterion is the relatively lengthy review process for submitted papers. Consequently, most of the concurrent validities or predictive validities over short periods of time may actually underestimate the true predictive validity of the GRE for this criterion.

Closely related to research productivity is the number of times a scientist's work is cited. Although there are clearly exceptions, higher quality work tends to receive more citations from colleagues. This measure will tend to capture the quality aspect of research, whereas research productivity is probably more strongly related to quantity. Creager (1966) reported a correlation of .30 between number of citations and research productivity.

Degree attainment, the successful completion of a degree program, and time to complete the degree are also sometimes used in validation studies. Although professional success does not necessarily require a degree, there are many important outcomes from degree completion, including legal constraints preventing full practice in a field for those without the degree. Degree attainment and time to complete are likely to be a function of many different performances ranging from scholastic to interpersonal, as well as events beyond the control of the student. The different performances are likely to be predicted by a number of individual-differences, only some of which are ability related. One potential problem with this criterion measure is that some doctoral programs give a terminal master's to students who leave the program early (willingly or unwillingly). Therefore, degree completion could be an imperfect measure of success. We attempted to include only research in which the criterion measured completion of the degree the student was admitted to pursue. Given that the link between GRE and degree attainment or time to complete is likely to be more distal and therefore less strong than the association between ability and all performances resulting in the degree attainment outcome, we expect small, though positive, correlations between GRE score and degree attainment.

Potential Moderators of GRE Validities

Several variables may moderate the relationship between scores on the GRE and performance in graduate school. First, the predictive validity of the GRE may vary by academic discipline. Although there are many similarities in some of the fundamental tasks required of all graduate students, there are differences in the type of training and demands of different academic areas. To investigate the impact of academic field on the predictive validity of the GRE tests, we conducted separate analyses for subsamples representing four broad disciplines: humanities, the social sciences, life sciences, and math–physical sciences.

A second potential moderator is whether or not English is a student's primary or preferred language. The validity of the GRE for non–native English speaking students is clearly a concern. With the test offered in English, non–native English speakers are often at a disadvantage. Research on native versus non–native English speakers suggests that at least some minimum level of proficiency in English is necessary for the GRE to be valid (Alderman, 1982). For non–native English speakers, we expect the GRE-Q to be a better predictor of graduate school performance because it is less dependent on verbal ability.

The final moderator examined was student age. Older students are likely to differ from more traditional students in work experience, time away from school, and family obligations. Despite these differences, it was expected that the GRE would be a comparably valid predictor for both younger and older students.

This article also examines the predictive validity of UGPA. Like GRE scores, UGPA is often used in the selection of graduate students. To provide a reference point for evaluating the GRE tests' predictive validities, we also conducted a meta-analysis of the validity of UGPA in predicting graduate school performance. Finally, we explored validities of combinations of the predictors examined in this meta-analysis.

In summary, this study meta-analytically addressed three main questions. First, to what extent is the GRE a valid predictor of graduate student performance? Second, is the GRE a better predictor for some criterion measures than others? Third, is the validity of the GRE moderated by academic discipline, native English speaking status, or age?

Method

The data collected from the studies were analyzed with the Hunter and Schmidt (1990) psychometric meta-analytic method. This method was preferred above others because it provides for estimating the amount of variance attributable to sampling error, range restriction, and unreliability. We used the artifact distributions described later to correct for the attenuating influences of artifacts on the observed correlations. The interactive meta-analysis procedure was used (Hunter & Schmidt, 1990, p. 165; Schmidt, Gast-Rosenberg, & Hunter, 1980). Data were analyzed with a program developed by Schmidt, Hunter, Viswesvaran, and colleagues with improvements that increased accuracy over the original Hunter and Schmidt (1990) method. These refinements included use of the mean observed correlation in the formula for sampling error variance and use of a nonlinear range restriction formula to estimate the standard deviation of corrected validities (Law, Schmidt, & Hunter, 1994a, 1994b).

Description of the Database

We gathered studies involving prediction of graduate school performance from several sources. To identify relevant research, we combined PsycLIT (1887–1999) and ERIC (1966–1999) searches with a search of *Dissertation Abstracts International* (1861–1998) and listings of ETS technical reports. The citation lists within all articles, dissertations, and technical reports were also examined to identify additional relevant studies. Each article was coded by one of the first two authors. The information collected from each article included the types of predictors, type of criterion, effect sizes, and sample sizes. Unreported effect sizes were computed from available information when possible. Information regarding moderators, range restriction, and criterion unreliability data was also recorded. Up to 39 different pieces of information for each bivariate relationship were coded.

To address potential overlap between samples across articles, dissertations, and technical reports, we identified studies with identical authors and evaluated their similarities. In articles with sample overlaps, the larger or more complete data were included in the meta-analysis, and the matching articles were excluded. When unclear, the authors of multiple studies were contacted to ensure that their samples were independent. When grades for a set of individual classes were reported as a criterion measure, correlations were averaged across courses. Finally, tests no longer offered by ETS (e.g., Profile tests) were not included in this meta-analysis.

Occasionally, a study presented only the results that were statistically significant. Studies that omitted any results based on significance tests were not included in the meta-analysis, because inclusion of results that are filtered out by significance tests could bias the findings (Hunter & Schmidt, 1990). Among the studies reporting validities for the GRE, less than 1% reported results screened by significance tests (i.e., where only significant correlations were reported and those nonsignificant were omitted).[1]

In some ETS technical reports, the reported correlations were aggregated across a large number of subsamples via one of two procedures: taking a

[1] If a study contained a multitude of predictors (some or all of the GRE-V, GRE-Q, GRE-A, and Subject Tests scores) with a number of different criteria (e.g., 1st-year grades, overall GPA, and ratings by faculty) but reported only those correlations between predictors and criteria that were significant, these studies were excluded from the database. Including only the significant and reported correlations would have inflated our meta-analytic estimates. We agree that the exclusion of studies presenting only significant findings is not entirely satisfactory. However, we believe that the biasing effects are likely to be far smaller than including the significant results and zero values for those effect sizes that were omitted by the study authors. When authors present only significant findings, we are left with four options. First, we can take the significant findings. Second, we can use methods developed by Hedges and Olkin to estimate what the missing effect sizes are likely to be. Third, we can include all nonreported nonsignificant findings as zero. Fourth, we can exclude such studies. We comment on each in turn.

Including only significant findings would clearly have resulted in an upward bias on estimated correlations (effect sizes) and was rejected. On the other hand, including all nonreported nonsignificant findings as zero would have downwardly biased the estimated correlations. This strategy was also rejected.

Although there are methods designed to include studies screened on the basis of significance (Hedges & Olkin, 1985), these methods were not used in this study. There were two main reasons for this. First, there were fewer than 10 studies that censored results based on significance, reporting only significant correlations. In this meta-analysis, our database contained 1,521 published and unpublished studies with 1,743 independent samples. As such, fewer than 1% of studies were excluded from our database as a consequence of filtering of reported results based on significance tests. Second, the statistical procedures described by Hedges and Olkin (1985) can provide precise estimates for the filtered studies (nonrandom sampling) only "when the censoring rule is known precisely" (p. 303). The approaches outlined by Hedges and Olkin (1985) require assumptions about how results were filtered. Unfortunately, few of the filtered studies reported what alpha level was used in the significance tests included in the screening. The extreme infrequency of studies with correlations omitted on the basis of significance in which the screening rule was known did not warrant their use in this meta-analysis.

median across studies or using the empirical Bayes method (Braun & Jones, 1985). These summaries could not be used in this study, because the sampling errors of the final correlation obtained with either method differ from the sampling error of the Pearson correlation. Including these aggregated and adjusted correlations would make the subsequent examinations of variance attributable to sampling error across studies inaccurate and uninterpretable. To address this problem, we contacted ETS researchers to obtain the data contributing to summary findings in the technical reports. ETS provided all available data on GRE validity to us in unaggregated form. We sorted these data, and incorporated them into our meta-analysis.

Although the reliability of coded meta-analytic data is high for meta-analyses, such as the current study, in which coding decisions are straightforward (Whetzel & McDaniel, 1988; Zakzanis, 1998), coder reliability was checked. The first two authors coded the same 16 randomly chosen articles near the beginning of the coding process. We computed percentage agreement between the two coders for the data that were relevant for the results presented in this study (sample sizes, correlations, standard deviations, reliability information, moderator variables, and variable types). Note that this resulted in a lower estimate of coder agreement because it eliminated coded information that was captured without error, such as journal names and author affiliation. This pruning resulted in 315 pieces of information for the agreement comparison. The authors agreed on 313 pieces of information, a rate of 99.4%. Both disagreements involved instances in which the sample size listed in one part of the results section did not correspond to sample information presented elsewhere in the paper being coded. Thus, the discrepancy between the two coders reflected unreliability in the published data. The decision on which sample size to include was largely arbitrary and involved less than a 100-person difference in sample sizes. The raw data from ETS were copied into the overall database via spreadsheet software and therefore were not subject to any coding errors. Overall, consistent with previous meta-analytic research (Whetzel & McDaniel, 1988; Zakzanis, 1998), coding errors were very infrequent and trivial in regard to their effect on the meta-analytic results.

The final database included 1,753 independent samples and 6,589 correlations across 82,659 graduate students. The correlations included relationships among eight criteria and five predictors. No analysis included multiple correlations from the same sample of individuals, and independence was not violated.

Range Restriction and Unreliability Artifact Distributions

When correcting for range restriction, great care must be taken to define the population of interest. In this study, all potential applicants to a graduate program were considered to be the population of interest. To correct for range restriction, the ratios of selected group standard deviations to applicant pool standard deviations (u values) are necessary. Complicating the issue, the standard deviations of scores for these groups have shifted across time. The GRE is scored back to a 1952 sample with a mean of 500 and a standard deviation of 100 (Briel, O'Neill, & Scheuneman, 1993). Over time, it appears that the standard deviation of GRE scores has increased. Current standard deviations of scores for graduate school applicants consistently exceed 100. To help address the problem with the changes in the standard deviation over time, we obtained the population standard deviations that were available in the GRE technical manuals and reports for the following years: 1952, 1967–1968, 1974–1976, 1988–1991, 1992–1995, and 1995–1996 (Briel et al., 1993; Conrad, Trisman, & Miller, 1977; ETS, 1996, 1997). Sample standard deviations from the meta-analysis were matched with applicant standard deviations closest to them in time.

In computations of range restriction values, standard deviations of samples reporting standard deviations were also linked with applicant standard deviations according to academic area. That is, the area listed in each individual study was matched with area groupings gathered by ETS before testing. The ETS area groupings are based on test taker self-reports of intended area of study. Standard deviations for these groups are published in the GRE technical manual. This matching was done because both mean scores and score standard deviations tend to differ by intended area of study. Failure to match according to area would be likely to result in overcorrection in that the total testing sample across areas is generally more variable (has larger standard deviations) than subareas. Artifact distribution information for all range restriction corrections is presented in Table 1.

Because the correlations of interest are between the tests and graduate school performance, the reliability of the measure of graduate school performance is an issue. The unreliability of performance measures, whether ratings, grades, or comprehensive exam scores, lowers the observed correlation between the performance measure and the GRE. Whenever possible, reliability estimates were used to correct for attenuation in validities for each criterion. Faculty ratings were corrected via a meta-analytically derived reliability for supervisor ratings of performance from the work performance literature (Viswesvaran, Ones, & Schmidt, 1996). The mean reliability in ratings was taken as .52. The reliability of grades was based on reliabilities from three studies of the reliability of college grades: Reilly and Warech (1993), Barritt (1966), and Bendig (1953). The internal consistency reliability values from these three studies were .84, .84, and .80, respectively. Artifact distribution information for all reliability corrections is presented in Table 1.

Results

We first present results for predictor–criterion combinations across academic areas of graduate programs. We then turn to an examination of validities for separate disciplines. For all meta-analyses, the average, sample-size/weighted correlation was computed across all studies (r_{obs}), as well as the standard deviation of observed correlations (SD_{obs}). The residual standard deviation of the correlations, after correction for statistical artifacts, was calculated next (SD_{res}). Finally, the operational validity coefficient (ρ) and the standard deviation for the true validities (SD_ρ) were computed, as well as the 90% lower credibility interval.

Overall Results Across Areas of Study

Meta-analyses of GRE and UGPA validities across disciplines were conducted separately for the following criteria: GGPA, 1st-year GGPA faculty ratings, comprehensive examination scores, degree attainment, time to degree completion, citation counts, and

Table 1
Artifact Distributions Used in the Meta-Analyses

Predictor or criterion	Mean U_{RR}	K_{RR}	Mean $R_{XX}^{1/2}$	K_{rel}
Predictor				
Verbal	.77	1,178	.96	9
Quantitative	.73	1,189	.95	9
Analytical	.74	1,032	.95	4
Subject	.82	67	.97	31
UGPA			.91	3
Criterion				
GGPA			.91	3
1st-year GGPA			.91	3
Faculty ratings			.73	1

Note. Mean U_{RR} = mean U ratio for range restriction; K_{RR} = number of ratios in the distribution; Mean $R_{XX}^{1/2}$ = mean of square root of the reliabilities; K_{rel} = number of reliabilities in the distribution; UGPA = undergraduate grade point average; GGPA = graduate grade point average.

research productivity. Corrections for range restriction and criterion unreliability were made when possible. Results by criterion type are presented in Table 2.

The meta-analytic results for predicting GGPA included studies reporting GPAs for 2 or more years of course work. The majority of studies used final GGPA. The meta-analytic results for predicting 1st-year GGPA included studies reporting GPAs based on one or two semesters of course work. The majority of studies had samples with a full year of grades.

The validities for GGPA were moderately large and nearly equal for the GRE-V (N = 14,156, k = 103), GRE-Q (N = 14,425, k = 103), GRE-A (N = 1,928, k = 20), and UGPA (N = 9,748, k = 58), with operational validities of .34, .32, .36, and .30, respectively. The standard deviations of these true validities (SD_ρ) were very small relative to meta-analyses of predictors of work performance (Hunter, 1983; Ones, Viswesvaran, & Schmidt, 1993; Pearlman et al., 1980).

The magnitude of the standard deviations of corrected validities is an indicator of the existence of moderators. The low standard deviations suggest that variables such as degree level (e.g., MA vs. PhD) and area of study are unlikely to meaningfully moderate the relationship between GRE-V, GRE-Q, GRE-A, and UGPA with GGPA. Also, the lower 90% credibility interval for each operational validity did not include zero, indicating that these three GRE scales and UGPA are valid for predicting GGPA across graduate departments, programs, and situations.

The Subject Tests had a larger operational validity (ρ = .41) that averaged .08 validity points higher than the other four predictors. The standard deviation of true validities for the Subject Tests was, on average, even lower than those for the other predictors, indicating that the validity of the Subject Tests are not likely to be affected by unexamined moderators.

Results for 1st-year GGPA were very similar to those for GGPA. The GRE-V (N = 46,615, k = 1,231), GRE-Q (N = 46,618, k = 1,231), GRE-A (N = 36,325, k = 1,080), and UGPA (N = 42,193, k = 1,178) validities were similar in magnitude. The standard deviations of operational validities were similar to those for GGPA. Again, the Subject Tests (ρ = .45) were found to be better predictors of 1st-year GGPA than the other four predictors by an average of .10 validity points.

The validities of GRE scores and UGPA for predicting comprehensive exam scores also are presented in Table 2. Operational validities for the GRE-V (N = 1,198, k = 11) and GRE-Q (N = 1,194, k = 11) were moderately large (.44 and .26, respectively), with standard deviations of true correlations very similar to those for GGPA. Lower 90% credibility intervals did not include zero, indicating validity generalization. UGPA (N = 592, k = 6) did not predict comprehensive scores nearly as well, with an operational validity of only .12. The lower 90% credibility interval for UGPA did not include zero for this criterion (the standard deviation of the true validities was estimated as zero). The GRE-V was found to be a somewhat better predictor of comprehensive exam scores than the GRE-Q, although the credibility intervals for the two predictors overlapped. Finally, the Subject Tests (N = 534, k = 4) were the best predictors of comprehensive exam scores (ρ = .51), exceeding the other predictors by an average of .24 correlation points.

GRE and UGPA correlations with faculty ratings are shown in Table 2. As described earlier, only faculty ratings of research ability, professional work, potential, or overall performance were included in this study. Operational validities for the GRE-V (N = 4,766, k = 35), GRE-Q (N = 5,112, k = 34), GRE-A (N = 1,982, k = 9), and UGPA (N = 3,695, k = 22) were similar. Much like the other criterion measures, the Subject Tests (N = 879, k = 12) had a larger correlation with faculty ratings, exceeding the other predictors by an average of .10. The operational validity of the Subject Tests was .50, whereas the corresponding validities were .42, .47, .35, and .35 for the GRE-V, GRE-Q, GRE-A, and UGPA, respectively. The standard deviations of operational validities were all small, leaving little room for moderators to operate.

Degree attainment in this study included studies that predicted graduation versus no graduation, success–failure, or staying in the graduate program versus dropping out. The GRE-V (N = 6,304, k = 32), GRE-Q (N = 6,304, k = 32), GRE-A (N = 1,233, k = 16), and UGPA (N = 6,315, k = 33) validities are presented in Table 2. Although uniformly positive, these validities ranged between .11 and .20 and were, on average, considerably smaller than those obtained for other criterion measures. Credibility intervals for all predictors included zero, except for the Subject Tests. For degree attainment, the operational validity of the Subject Tests (ρ = .39, N = 2,575, k = 11) was again larger than all other predictors. The standard deviations of the operational validities were noticeably larger than those for the GPA criteria, comprehensive exam scores, or faculty ratings and ranged from .16 to .30. This suggests that the relationship between the predictors and degree attainment may be moderated by other variables. For example, the differential base rates of graduation from programs may affect the size of the relationship between GRE test scores and degree attainment.

Relatively few studies examined the validity of the GRE scales and UGPA for predicting the amount of time it takes students to complete degrees. As shown in Table 2, operational validities were small and varied in direction. Moderate and zero correlations (.28 and .02, respectively) were obtained for the GRE-V (N = 160, k = 3) and the Subject Tests (N = 66, k = 2) for time to completion. Small negative correlations (−.12 and −.08, respectively) were obtained for the GRE-Q (N = 160, k = 3) and UGPA (N = 629, k = 5).

The criterion, research productivity, includes studies that used measures of research productivity, distinguished between students with publications and those without, included number of publications, or noted number of submissions to journals or number of conference papers presented. Data were available only for the GRE-V (N = 3,328, k = 18), GRE-Q (N = 3,328, k = 18), and Subject Tests (N = 3,058, k = 16). Although uniformly positive, the 90% credibility intervals included zero for this criterion except for the Subject Tests, which were moderately correlated with research productivity (.21). It should be noted that the majority of these data came from Creager (1966), who examined National Science Foundation Fellowship applicants.

Citation count was the final criterion examined. Moderate correlations were obtained for the GRE-V (N = 2,306, k = 12), GRE-Q (N = 2,306, k = 12), and Subject Tests (N = 2,306, k = 12). Credibility intervals for all three predictors did not include zero. The GRE-Q and the Subject Tests were similarly correlated (.23 and .24, respectively) with citation counts, with a somewhat smaller validity for the GRE-V (.17). All of the data examined for this criterion were from Creager (1966).

Table 2
Meta-Analysis of GRE and UGPA Validities: Total Sample

Predictor	N	k	r_{obs}	SD_{obs}	SD_{res}	ρ	SD_ρ	90% credibility interval
				GGPA				
Verbal	14,156	103	.23	.14	.10	**.34**	.15	.09 to .59
Quantitative	14,425	103	.21	.11	.06	**.32**	.08	.19 to .45
Analytical	1,928	20	.24	.12	.04	**.36**	.06	.26 to .46
Subject	2,413	22	.31	.12	.05	**.41**	.07	.30 to .52
UGPA[a]	9,748	58	.28	.13	.10	**.30**	.11	.12 to .48
				1st-year GGPA				
Verbal	45,615	1,231	.24	.19	.09	**.34**	.12	.14 to .54
Quantitative	45,618	1,231	.24	.19	.08	**.38**	.12	.18 to .58
Analytical	36,325	1,080	.24	.19	.06	**.36**	.09	.21 to .51
Subject	10,225	98	.34	.11	.03	**.45**	.04	.38 to .52
UGPA[a]	42,193	1,178	.30	.18	.10	**.33**	.10	.17 to .49
				Comprehensive exam scores[b]				
Verbal[c]	1,198	11	.34	.16	.12	**.44**	.15	.19 to .69
Quantitative[c]	1,194	11	.19	.11	.04	**.26**	.06	.16 to .36
Subject[d]	534	4	.43	.07	.00	**.51**	.00	.51 to .51
UGPA[a]	592	6	.12	.05	.00	**.12**	.00	.12 to .12
				Faculty ratings				
Verbal	4,766	35	.23	.12	.08	**.42**	.14	.19 to .65
Quantitative	5,112	34	.25	.10	.02	**.47**	.04	.40 to .54
Analytical	1,982	9	.23	.05	.00	**.35**	.00	.35 to .35
Subject	879	12	.30	.16	.11	**.50**	.18	.20 to .80
UGPA[a]	3,695	22	.25	.12	.10	**.35**	.14	.12 to .58
				Degree attainment[a]				
Verbal	6,304	32	.14	.14	.12	**.18**	.16	−.08 to .44
Quantitative	6,304	32	.14	.17	.15	**.20**	.20	−.13 to .53
Analytical	1,233	16	.08	.25	.22	**.11**	.30	−.38 to .60
Subject	2,575	11	.32	.16	.14	**.39**	.17	.11 to .67
UGPA[a]	6,315	33	.12	.17	.16	**.12**	.16	−.14 to .38
				Time to complete[b,e]				
Verbal	160	3	.21	.07	.00	**.28**	.00	.28 to .28
Quantitative	160	3	−.08	.05	.00	**−.12**	.00	−.12 to −.12
Subject	66	2	.02	.05	.00	**.02**	.00	.02 to .02
UGPA[a]	629	5	−.08	.10	.04	**−.08**	.04	−.15 to −.01
				Research productivity[b]				
Verbal	3,328	18	.07	.12	.10	**.09**	.13	−.12 to .30
Quantitative	3,328	18	.08	.10	.07	**.11**	.09	−.04 to .26
Subject	3,058	16	.17	.13	.10	**.21**	.12	.01 to .41
				Publication citation count[b,f]				
Verbal	2,306	12	.13	.09	.05	**.17**	.06	.07 to .27
Quantitative	2,306	12	.17	.09	.04	**.23**	.05	.15 to .31
Subject	2,306	12	.20	.09	.03	**.24**	.04	.17 to .31

Note. GRE = Graduate Record Examinations; UGPA = undergraduate grade point average; k = number of studies; r_{obs} = sample-size-weighted average correlation; SD_{obs} = standard deviation of observed correlations; SD_{res} = residual standard deviation; ρ = estimated operational validity; SD_ρ = standard deviation of true validity correlations; GGPA = graduate grade point average. Boldface entries indicate best estimates of predictor validity.
[a] Not corrected for range restriction. [b] Not corrected for criterion unreliability. [c] Most study comprehensive exam scores with Verbal and Quantitative samples are from the social sciences (k = 11). [d] Comprehensive exam scores with Subject Tests samples are from the social sciences. [e] All time to complete studies data from the social sciences. [f] All studies from Creager (1966).

Results for Different Areas of Study

To examine whether GRE and UGPA validities for different disciplines differed from the results across areas, we separated studies with samples from four broad discipline groups from the total sample and meta-analyzed. The four subareas were humanities, social sciences, life sciences, and math–physical sciences. The fields listed in the studies included in the humanities group were art, music, English, literature, liberal arts, philosophy, foreign language, humanities, and speech. The fields represented in the social sciences samples were psychology, education, history, social science, business, sociology, economics, social work, anthropology, political science, occupational therapy, library science, and public administration. The specific fields in the life sciences group were biology, nursing, agriculture, veterinary medicine, natural sciences, and forestry. Finally, the math–physical sciences group included mathematics, physics, chemistry, computer science, geosciences, geology, statistics, engineering, and math–physical sciences. Results for these subsamples, presented in Tables 3–6, should be interpreted with caution, as smaller sample sizes compared to the overall analyses result in greater sampling error and less stable estimates.

Notably, the results for the separate discipline groups were highly similar to those for the overall sample. The GRE-V, GRE-Q, GRE-A, and UGPA operational validities were very similar for GGPA, 1st-year GGPA, and faculty ratings. The Subject Tests were consistently the best predictor within subgroups, with the exception of degree attainment. Finally, the standard deviations of true validities were also small across analyses, much like the overall sample. Note that many of these subarea analyses were based on relatively small sample sizes and should be interpreted with caution.

Results for Non–Native English Speakers and Nontraditional Students

The studies examining the validity of the GRE for non–native English speakers were not included in the overall analysis and were considered separately. A meta-analysis of these studies is presented in Table 7 for GGPA and 1st-year GGPA.

For GGPA, the operational validity (ρ = .36) of the GRE-V (N = 1,764, k = 6) was quite similar for native and non–native English speakers, whereas the operational validity of the GRE-Q (N = 1,705, k = 5) was larger (ρ = .53) than that observed in the overall sample. The operational validities for the GRE-V, GRE-Q, and GRE-A in predicting 1st-year GGPA were similar to the overall, native English speaking samples with operational validities of .22, .40, and .35, respectively. The standard deviations of the operational validities for the non–native English speaking sample tended to be even smaller than those for the total sample. This was partially due to the small number of validities contributing to these analyses.

Table 3
Meta-Analysis of GRE and UGPA Validities for Prediction of GGPA: Subdisciplines

Subdiscipline	N	k	r_{obs}	SD_{obs}	SD_{res}	ρ	SD_ρ	90% credibility interval
Humanities								
Verbal	999	12	.22	.22	.19	**.32**	.27	−.12 to .76
Quantitative	999	12	.18	.14	.07	**.27**	.11	.09 to .45
Analytical	63	2	.33	.09	.00	**.48**	.00	.48 to .48
Subject	128	3	.37	.27	.22	**.49**	.29	.01 to .97
UGPA[a]	63	2	.13	.16	.00	**.14**	.00	.14 to .14
Social science								
Verbal	7,610	55	.27	.13	.07	**.39**	.11	.21 to .57
Quantitative	7,260	54	.23	.11	.03	**.34**	.04	.27 to .41
Analytical	957	9	.26	.14	.08	**.38**	.12	.15 to .58
Subject	1,857	12	.30	.11	.05	**.40**	.06	.30 to .50
UGPA[a]	4,132	32	.29	.11	.07	**.32**	.07	.21 to .43
Life science								
Verbal	1,563	11	.27	.09	.00	**.39**	.00	.39 to .39
Quantitative	1,563	11	.24	.08	.00	**.37**	.00	.37 to .37
Analytical	479	4	.24	.08	.00	**.36**	.00	.36 to .36
Subject	84	2	.31	.04	.00	**.42**	.00	.42 to .42
UGPA[a]	1,947	10	.26	.11	.08	**.28**	.09	.13 to .43
Math–physical science								
Verbal	827	12	.21	.18	.13	**.30**	.19	−.01 to .61
Quantitative	827	12	.25	.15	.06	**.38**	.10	.22 to .54
Analytical	201	4	.24	.15	.04	**.36**	.06	.26 to .46
Subject	95	3	.30	.15	.00	**.40**	.00	.40 to .40
UGPA[a]	252	5	.38	.08	.00	**.41**	.00	.41 to .41

Note. GRE = Graduate Record Examinations; UGPA = undergraduate grade point average; GGPA = graduate grade point average; k = number of studies; r_{obs} = sample-size-weighted average correlation; SD_{obs} = standard deviation of observed correlations; SD_{res} = residual standard deviation; ρ = estimated operational validity; SD_ρ = standard deviation of true validity correlations. Boldface entries indicate best estimates of predictor validity.
[a] Not corrected for range restriction.

Table 4
Meta-Analysis of GRE and UGPA Validities for Prediction of 1st-Year GGPA: Subdisciplines

Subdiscipline	N	k	r_{obs}	SD_{obs}	SD_{res}	ρ	SD_ρ	90% credibility interval
Humanities								
Verbal	6,152	180	.28	.18	.04	**.40**	.06	.30 to .50
Quantitative	6,152	180	.23	.20	.09	**.35**	.13	.14 to .56
Analytical	4,277	143	.22	.20	.08	**.33**	.12	.13 to .53
Subject	1,317	24	.32	.14	.05	**.42**	.06	.32 to .52
UGPA[a]	5,489	167	.30	.18	.09	**.33**	.09	.18 to .48
Social science								
Verbal	22,375	486	.26	.17	.08	**.37**	.11	.19 to .55
Quantitative	22,378	486	.24	.18	.09	**.37**	.13	.16 to .58
Analytical	17,917	433	.26	.17	.06	**.38**	.10	.22 to .54
Subject	5,081	34	.36	.09	.00	**.47**	.00	.47 to .47
UGPA[a]	20,547	468	.30	.16	.09	**.33**	.09	.18 to .48
Life science								
Verbal	8,616	233	.24	.18	.07	**.34**	.10	.18 to .50
Quantitative	8,616	233	.23	.17	.01	**.35**	.02	.32 to .38
Analytical	7,762	208	.22	.17	.03	**.34**	.04	.27 to .41
Subject	852	13	.25	.12	.00	**.33**	.00	.33 to .33
UGPA[a]	8,446	225	.31	.19	.12	**.34**	.13	.13 to .55
Math–physical science								
Verbal	8,076	329	.16	.23	.10	**.24**	.15	−.01 to .49
Quantitative	8,076	329	.25	.22	.08	**.37**	.11	.19 to .55
Analytical	6,333	295	.22	.22	.05	**.33**	.07	.22 to .44
Subject	2,621	25	.35	.11	.03	**.47**	.04	.40 to .54
UGPA[a]	7,288	315	.31	.22	.10	**.34**	.11	.16 to .52

Note. GRE = Graduate Record Examinations; UGPA = undergraduate grade point average; GGPA = graduate grade point average; k = number of studies; r_{obs} = sample-size-weighted average correlation; SD_{obs} = standard deviation of observed correlations; SD_{res} = residual standard deviation; ρ = estimated operational validity; SD_ρ = standard deviation of true validity correlations. Boldface entries indicate best estimates of predictor validity.
[a] Not corrected for range restriction.

Nontraditional students have also been examined in only a few studies. A meta-analysis of these studies is also presented in Table 7 for GGPA. The two studies in Table 7 involved students who were more than 30 years old. Although samples sizes were not large, the correlations were positive across the samples. The GRE appears to be a valid predictor of GGPA and 1st-year GGPA for older students.

A final moderator that could be of some general concern is the effect of grade inflation over time on the validity of the GRE for predicting graduate school grades. If grade inflation has reduced

Table 5
Meta-Analysis of GRE and UGPA Validities for Prediction of Faculty Ratings: Subdisciplines

Subdiscipline	N	k	r_{obs}	SD_{obs}	SD_{res}	ρ	SD_ρ	90% credibility interval
Humanities								
Verbal	311	4	.41	.15	.08	**.72**	.13	.51 to .93
Quantitative	311	4	.31	.10	.00	**.58**	.00	.58 to .58
Social science								
Verbal	1,965	19	.20	.13	.07	**.37**	.13	.16 to .58
Quantitative	1,965	19	.20	.13	.07	**.38**	.13	.17 to .59
Analytical	941	6	.20	.04	.00	**.37**	.00	.37 to .37
Subject	515	8	.23	.15	.08	**.38**	.14	.15 to .61
UGPA[a]	1,132	14	.19	.16	.12	**.27**	.17	−.01 to .55
Life science								
Verbal	854	4	.23	.08	.00	**.42**	.00	.42 to .42
Quantitative	854	4	.22	.05	.00	**.41**	.00	.41 to .41
UGPA[a]	836	3	.25	.09	.07	**.34**	.10	.18 to .50
Math–physical science								
Verbal	508	4	.23	.11	.04	**.42**	.07	.31 to .53
Quantitative	508	4	.34	.04	.00	**.63**	.00	.63 to .63

Note. GRE = Graduate Record Examinations; UGPA = undergraduate grade point average; k = number of studies; r_{obs} = sample-size-weighted average correlation; SD_{obs} = standard deviation of observed correlations; SD_{res} = residual standard deviation; ρ = estimated operational validity; SD_ρ = standard deviation of true validity correlations. Boldface entries indicate best estimates of predictor validity.
[a] Not corrected for range restriction.

Table 6
Meta-Analysis of GRE and UGPA Validities for Prediction of Degree Attainment: Subdisciplines

Subdiscipline	N	k	r_{obs}	SD_{obs}	SD_{res}	ρ	SD_ρ	90% credibility interval
Humanities								
Verbal	61	2	.41	.15	.08	**.72**	.13	.51 to .93
Quantitative	61	2	.12	.20	.09	**.17**	.12	−.03 to .37
Analytical	61	2	.12	.16	.00	**.16**	.00	.16 to .16
UGPA[a]	61	2	−.02	.03	.00	**−.02**	.00	−.02 to −.02
Social science								
Verbal	2,062	14	.17	.17	.14	**.22**	.18	−.08 to .52
Quantitative	2,062	14	.22	.15	.11	**.31**	.15	.06 to .56
Analytical	334	5	.37	.24	.20	**.49**	.26	.06 to .92
Subject	1,022	6	.24	.10	.06	**.30**	.07	.19 to .41
UGPA[a]	2,077	15	.14	.23	.21	**.14**	.21	−.20 to .48
Life science								
Verbal	1,055	6	.03	.06	.00	**.03**	.00	.03 to .03
Quantitative	1,055	6	−.07	.08	.01	**−.09**	.01	−.11 to −.07
Analytical	644	5	−.07	.13	.09	**−.10**	.12	−.30 to .10
UGPA[a]	1,051	6	.05	.09	.05	**.05**	.05	−.03 to .13
Math–physical science								
Verbal	1,747	9	.20	.16	.13	**.26**	.17	−.02 to .54
Quantitative	1,747	9	.22	.16	.13	**.31**	.18	.01 to .61
Analytical	194	4	.10	.03	.00	**.14**	.00	.14 to .14
UGPA[a]	1,747	9	.22	.13	.12	**.22**	.11	.04 to .40

Note. GRE = Graduate Record Examinations; UGPA = undergraduate grade point average; k = number of studies; r_{obs} = sample-size-weighted average correlation; SD_{obs} = standard deviation of observed correlations; SD_{res} = residual standard deviation; ρ = estimated operational validity; SD_ρ = standard deviation of true validity correlations. Boldface entries indicate best estimates of predictor validity.
[a] Not corrected for range restriction.

the information in grades, we would expect smaller validities in those samples with inflated grades and larger validities in those samples without inflated grades. Given that grade inflation has increased over time, in examining this moderator, we used year of study as an indirect measure of grade inflation. We correlated year of study with the observed correlations of GRE-V and GRE-Q with 1st-year GGPA. We found no relationship between year of study and observed validity of the GRE, with correlations of −.006 and −.007 between year of study and observed GRE-V (N = 1,231) and GRE-Q (N = 1,213) correlations, respectively.

Table 7
Meta-Analysis of GRE and UGPA Validities for Non–Native English Speaking and Nontraditional Graduate Students

Criterion	N	k	r_{obs}	SD_{obs}	SD_{res}	ρ	SD_ρ	90% credibility interval
			Non–native English speaking students					
GGPA								
Verbal	1,764	6	.25	.07	.00	**.36**	.00	.36 to .36
Quantitative	1,705	5	.37	.08	.00	**.53**	.00	.53 to .53
1st-year GGPA								
Verbal	6,855	360	.15	.27	.14	**.22**	.20	−.11 to .55
Quantitative	6,796	359	.27	.25	.11	**.40**	.16	.14 to .66
Analytical	6,777	358	.23	.26	.12	**.35**	.17	.07 to .63
Faculty ratings								
Verbal	190	2	.40	.11	.00	**.70**	.00	.70 to .70
Quantitative	190	2	.41	.13	.00	**.74**	.00	.74 to .74
			Nontraditional students					
GGPA								
Verbal	953	2	.34	.05	.00	**.45**	.00	.45 to .45
Quantitative	953	2	.23	.03	.00	**.31**	.00	.31 to .31

Note. GRE = Graduate Record Examinations; UGPA = undergraduate grade point average; GGPA = graduate grade point average; k = number of studies; r_{obs} = sample-size-weighted average correlation; SD_{obs} = standard deviation of observed correlations; SD_{res} = residual standard deviation; ρ = estimated operational validity; SD_ρ = standard deviation of true validity correlations. Boldface entries indicate best estimates of predictor validity.

Validities for Combinations of Predictors

Meta-analysis can be used to estimate the validity of combinations of predictors. Meta-analytically derived matrices of intercorrelations among predictors and intercorrelations among criterion measures are used to estimate the validity of composites of predictors (Viswesvaran & Ones, 1995).

The methodology of re-creating intercorrelation matrices based on meta-analytically derived estimates has been used in several previous studies (e.g., Hom, Caranikas-Walker, Prussia, & Griffeth, 1992; Peters, Hartke, & Pohlmann, 1985; Premack & Hunter, 1988). The expectation in creating a meta-analytically derived matrix of intercorrelations is that it represents population-level relationships more accurately than can any given study. Meta-analytically constructed intercorrelation matrices have been used in regression analyses (e.g., Ones et al., 1993; Schmidt & Hunter, 1998) as well as in structural equation modeling (e.g., Hom et al., 1992; Premack & Hunter, 1988; Verhaeghen & Salthouse, 1997). A number of methodological articles have discussed the pros and cons of meta-analytically derived matrices as input for further statistical analyses and have highlighted some potential problems as well as solutions (e.g., Becker & Schram, 1994; Shadish, 1996; Viswesvaran & Ones, 1995). Of the potential problems identified, four are relevant to our meta-analytically derived matrix of intercorrelations. We discuss these problems in turn.

First, there is the issue of missing data or small amounts of data for some of the cells of the matrix, contributing to higher levels of imprecision in some analyses. In this study, we were careful not to include any variables in our matrix for which intercorrelations were not available for some of the cells. That is, no analyses of multiple predictors were conducted that required information from predictors or criteria that resulted in missing cells. Furthermore, sample sizes for all cells in our matrix were moderate to large. The smallest sample in the matrix was 592 students based on six studies. Nevertheless, of course, analyses with smaller samples should be regarded with more caution than those with larger sample sizes.

The second issue is mistakenly including studies that appear to measure the same constructs when they do not. This is a general issue in all meta-analyses and one that is not likely to be a problem in this study for two reasons. First, all of the predictors were measures that either have been equated (GRE scales) or are very similar (high school grades). Second, we carefully separated the criterion measures into different groups to avoid the mistake of placing all measures into a "graduate school performance" category.

The third issue involves the homogeneity of the meta-analytic correlations in the intercorrelation matrix. If different populations have been combined and these populations have different correlations, then the analyses using the overall matrix could be questioned. Another way to state this potential problem is to say that all corrected correlations within the matrix should be positive and have small associated standard deviations. If the estimates in the matrix deviate from this, then there may be problems with the matrix, especially when it is used to answer multivariate questions. This is a legitimate concern. Heterogeneity may be present if there is residual variability after accounting for variation due to sampling error, dichotomization, differences in unreliability, and differences in range restriction across studies. That is, if there is some residual variation present that is not accounted for by sampling error, by interstudy range restriction differences or interstudy criterion unreliability differences in the predictor validities, or by intercorrelations between the predictors, this could be due to true heterogeneity in correlations or other sources of artifactual variation (computational errors, transcription errors, and construct validity problems) that cannot be addressed. Under such a scenario, one cannot be certain of the source of the remaining variation.

One nonartifactual source of variation that may be present is the effect of compensatory selection.[2] Presence of compensatory selection would primarily influence the intercorrelations between GRE scores and UGPA observed in different studies. This may occur if the selection decision process differs across universities. Specifically, if compensatory selection is used to varying degrees across universities, then the predictor intercorrelations from the student sample may be both attenuated and heterogeneous to an unknown extent from the population intercorrelations. Therefore, analyses involving the combination of multiple predictors, including UGPA, should be scrutinized carefully for heterogeneity and should be regarded with some caution if substantial heterogeneity is found.

Two important facts can address this potentially significant question in our meta-analysis. First, we should note that the intercorrelations among the GRE scales were based on population intercorrelations taken from the test manuals and are not the result of compensatory selection. These intercorrelations are unlikely to have heterogeneity problems, and the small residual standard deviations support this position. Only the correlations between UGPA and GRE scores may be subject to the just-discussed "compensatory" selection effect. Second, and most important, for the correlations between UGPA and GRE tests, the residual variations after accounting for sampling error and other artifacts were relatively small, making heterogeneity less of a concern. Residual standard deviations after accounting for sampling error and range restriction ranged from .02 to .06. Little variation remained, suggesting that heterogeneity problems are not likely. This provides indirect evidence that compensatory selection was not operating to a great extent in the studies that contributed to our database. Nonetheless, researchers should consistently report intercorrela-

[2] As discussed by Dawes (1971, 1975), compensatory selection can result in attenuated or negative correlations between predictors. Compensatory selection occurs when higher scores on one predictor are allowed to compensate for lower scores on other predictors in admitting students. Both school selectivity and admission policy can affect these relationships. Selectivity tends to affect the extent to which test scores and UGPA values can differ (students with extremely low UGPA or GRE scores are unlikely to be admitted). Policy, in the form of cutoffs or multiple hurdle procedures, can also affect the relationship between the predictors. To completely address this problem, a multivariate approach would need to be adopted. Differences in compensatory selection across schools would produce larger observed variances than what we would anticipate as a result of sampling error and other statistical artifacts. The primary effect on our results would be a decrease in the correlation between UGPA and GRE scores and a large SD_ρ associated with this correlation. If there is not a large SD_ρ we can have increased confidence that compensatory selection is not obscuring the true relationship between UGPA and the other predictors.

tions among predictors so that the issue of compensatory selection can be more directly studied in future meta-analyses.

The fourth issue is artifactual variation influencing some or all of the correlations in the matrix. In this case, differences in unreliability and range restriction across studies will result in a matrix that does not represent the population matrix. In this study, we corrected for both unreliability and range restriction using the Hunter and Schmidt (1990) artifact distribution method. If the assumptions for the artifact distribution meta-analytic method hold, this is unlikely to be a concern in our study.

To examine the validity of combinations of predictors, we used meta-analysis to estimate a matrix of intercorrelations between predictors as well as intercorrelations between criterion measures (Viswesvaran & Ones, 1995). All of the studies included in this meta-analysis were examined, and intercorrelations between predictors or criteria were coded. Each cell in the matrix was treated as a separate meta-analysis. The resulting matrix of intercorrelations (shown in Table 8) was used to compute unit-weighted composites of predictors. Nunnally (1978) provided the following equation for the correlation between two unit-weighted composites: $r_{wy} = \Sigma R_{wy}/\Sigma R_w^{1/2}\Sigma R_y^{1/2}$, where ΣR_{wy} is the sum of the predictive validities, ΣR_w is the sum of the matrix of predictor intercorrelations, and ΣR_y is the sum of the matrix of criterion intercorrelations. Unit-weighted composites are presented in Table 9.

These values estimate the validity of the combined predictors in predicting a unit-weighted composite of GGPA and faculty ratings. These two criteria were combined in this analysis for two reasons. First, reasonably large samples were available for both measures. Second, we felt they represented two important aspects of graduate student performance. The composites tend to improve prediction of graduate student success. Note, however, that the GRE Subject Tests alone predicted the criterion composite nearly as well as, and in some cases better than, the composites. The addition of the Subject Tests to any composite leads to a noticeable improvement in prediction. For students who had not taken the GRE Subject Tests in their area, the combination of GRE-V, GRE-Q, and UGPA produced a validity of .53, a substantial operational validity that exceeded that of the Subject Tests alone (.49). Only unit-weighted composites were estimated for two reasons. First, unit weights are a robust method for combining information, especially with predictors that are positively intercorrelated and similar in predictive validity. Second, if the meta-analytically derived predictor intercorrelations are not precise estimates as a result of compensatory selection, optimal weights derived from this matrix would not be useful. If one assumes that the intercorrelation between UGPA and other predictors has been suppressed by compensatory selection, the composites could be easily recomputed with Nunnally's (1978) equation using a larger intercorrelation. In most cases, the effect of this on the results in Table 9 would be that UGPA yields less incremental validity.

Discussion

The GRE-V, GRE-Q, GRE-A, and Subject Tests were found to be generalizably valid predictors of GGPA, 1st-year GGPA, faculty ratings, comprehensive examination scores, citation counts, and, to a lesser extent, degree attainment. The very small corrected standard deviations of the validity distributions suggest that the validity of the GRE generalizes across areas, departments, and situations and is not likely to be strongly moderated by unexamined variables. The GRE Subject Tests were consistently better predictors of all criteria (except for time to completion, which was not predicted by any measure) and were generalizably valid predictors of research productivity. GRE-V, GRE-Q, GRE-A, and UGPA validities were very similar to each other across the multiple criteria examined in this research: GGPA, 1st-year GGPA, faculty ratings, and comprehensive exams.

All studies included in this meta-analysis can be considered quasi-predictive studies. Concurrent studies of the GRE's relationship with graduate school performance were not conducted. GRE scores were submitted to graduate schools as part of the admission process. Thus, across studies, it is possible that some contamination of criterion measures exists and may have influenced the results. This might take the form of faculty being aware of a student's GRE scores and, in turn, this knowledge influencing grades, ratings, or other outcomes. The influence of this type of contamination is unlikely. To the best of our knowledge, GRE scores were not made available to faculty to aid in their grading or ratings of graduate students, and faculty rarely refer to them after admissions. This reduces the possibility and magnitude of contamination. Finally, some of a student's work in graduate school

Table 8
Meta-Analytically Derived Matrix of GRE, UGPA, and Criterion Measure Intercorrelations

Variable	1	2	3	4	5	6	7	8
1. GGPA	.91	.76		.30	.34	.32	.36	.41
2. Ratings	1,575 (7)	.54		.35	.42	.47	.35	.50
3. Comps				.12	.44	.26		.51
4. UGPA	9,748 (58)	3,695 (22)	592 (6)	.91	.24	.18	.24	.20
5. Verbal	14,156 (103)	4,766 (35)	1,198 (11)	6,897 (23)	.96	.56	.77	.62
6. Quantitative	14,425 (103)	5,122 (34)	1,194 (11)	6,897 (23)	145,912 (7)	.95	.73	.55
7. Analytical	1,928 (20)	1,982 (9)		3,888 (8)	3,895 (2)	3,895 (2)	.95	.52
8. Subject	2,413 (22)	879 (12)	534 (4)	892 (7)	78,728 (34)	78,728 (34)	31,025 (16)	.82

Note. Estimated correlations and standard deviations of true validity correlations are presented above the diagonal, reliabilities are presented on the diagonal, and sample sizes are presented below the diagonal. Values outside of parentheses are sample sizes, and values within parentheses are number of studies. Correlations with GRE scales have been corrected for range restriction. Correlations with GGPA and ratings have been corrected for unreliability. GRE = Graduate Record Examinations; UGPA = undergraduate grade point average; GGPA = graduate grade point average; Ratings = faculty ratings; Comps = comprehensive examination scores.

Table 9
GRE and UGPA Unit-Weighted Composite Predicting GGPA and Faculty Ratings

Predictor set	Predictive validity of unit-weighted composite	Predictive validity of composite plus UGPA (unit weighted)
Verbal	.41	.48
Quantitative	.42	.50
Analytical	.38	.46
Subject	.49	.54
Verbal + Quantitative	.46	.53
Verbal + Quantitative + Analytical	.45	.50
Verbal + Quantitative + Subject	.52	.56
Verbal + Quantitative + Analytical + Subject	.50	.54

Note. GRE = Graduate Record Examinations; UGPA = undergraduate grade point average; GGPA = graduate grade point average.

occurs with faculty outside the student's program or area, and these faculty are unlikely to have had any contact with a student's GRE scores.

Our results were quite consistent with previous personnel psychology research on the relationships among job knowledge, general cognitive ability, and work performance (Borman et al., 1993; Schmidt et al., 1986). Previous research indicates that general ability measures are predictive of performance on all jobs, with general ability exerting its primary influence indirectly through job knowledge (Borman et al., 1993; Schmidt et al., 1986). Individuals invest their ability and time in the acquisition of declarative and procedural knowledge. Work performance is then a function of a person's acquired declarative knowledge and skill when the individual chooses or is motivated to use them (McCloy et al., 1994). Therefore, general mental ability is a more distal predictor of performance, because job performance is a direct function of invested general ability in the form of declarative and procedural knowledge. The results of our meta-analysis fit this theoretical model and mirror previous findings.

First, consistent with previous meta-analyses of general cognitive ability (Hunter, 1980), the GRE-V, GRE-Q, and GRE-A, also general cognitive ability measures, predicted subsequent graduate student performance across all examined disciplines. Second, the Subject Tests, a more proximal determinant of graduate school performance, were better predictors of graduate school success and were quite similar to job knowledge tests in terms of their predictive power. Finally, general mental ability added little incremental validity when added to the Subject Tests. This too is consistent with previous research on job knowledge and work performance, in which the direct path between general mental ability and work performance has been shown to be smaller than the indirect path through job knowledge (Borman et al., 1993; Schmidt et al., 1986).

A major difference between these results and previous research on work performance is the magnitude of the difference in predictive validities between the general ability measures and the job knowledge measures. The General Tests were somewhat weaker predictors than the Subject Tests, more so than is typically found between general cognitive ability and job knowledge measures (Schmidt & Hunter, 1998). This difference cannot be attributed to reliability differences, because the Subject Tests are, in fact, slightly less reliable than the General Tests (Briel et al., 1993).

One explanation is that the criterion of school performance is more heavily determined by declarative knowledge than job performance in most work settings. Few occupations involve direct and comprehensive tests of job knowledge. Some aspects of graduate school performance, such as comprehensive examination scores, are themselves job knowledge tests.

An alternative explanation is that performance on the Subject Tests may be determined by actual knowledge acquired and ability along with motivation and interest. That is, subject tests could also be measuring interests and, therefore, subsequent motivation to study and master a field. Because much of graduate school involves investing a great deal of time in learning new material, this could explain the interest and motivation effect. Setting knowledge aside, a student who likes and studies a subject such as psychology from the start of college may be a more interested and motivated student than one who decides to switch fields. Presumably, the Subject Tests would reflect interest as well as knowledge as the two are intertwined. However, this hypothesis would need to be tested, because it is not unreasonable to assume that all job knowledge measures also measure interest. For example, one would presume that an individual's mechanical knowledge is partially a function of interest in working with machines. All job knowledge requires some motivation to invest the time and cognitive resources in acquiring declarative and procedural knowledge.

Supporting the interest hypothesis are the low correlations for degree attainment with the exception of the Subject Tests. Although uniformly positive, all measures but the Subject Tests had credibility intervals that included zero. These results are not surprising, because there are a large number of noncognitive and situational variables that can strongly affect degree attainment. The positive results obtained in this meta-analysis are promising and suggest that selection based on the GRE certainly does not result in students who are less likely to complete their programs and is likely to actually help select successful students. This is especially the case for the Subject Tests, which had the largest correlations. The large correlation between the Subject Tests and degree attainment, as well as other criteria, may in part be due to interest in a subject area. This interest could result in persistence and completion of graduate school, although the effect could simply be a function of job knowledge facilitating completion. In other words, those with high scores on the Subject Tests may be no more interested or motivated but simply may have a head start on their classmates. Again, additional study of this area is needed to answer this question fully and to disentangle the effects.

The only issue surrounding the use of the Subject Tests is that they may not be a very good indicator for those people who have not spent any time learning about a particular subject. In these cases, the prior research on cognitive ability and job knowledge suggests that scores on the GRE-V and GRE-Q would predict an individual's later acquisition of subject knowledge. Hence, the value of the General Tests is for those whose undergraduate degrees are in an area other than the one they apply for in graduate school.

Although our results indicate that the GRE has a valuable place in graduate student selection, there remains much room to increase the validity of our graduate student selection systems. These improvements can be made with additional predictors or improved data combination methods.

For other predictors to provide incremental validity, they must be correlated with the criterion and typically weakly related, or ideally uncorrelated, with other predictors used in the selection system. Determinants of graduate school performance that are related to interest, independence, and motivation are likely to fulfill this role. Operationalizations of personality and interest measures can take many different forms. Currently, this information is gathered from personal statements and letters of recommendation, but it could be more systematically and reliably collected. Personality and interest measures generally exhibit weak correlations with cognitive ability measures (Ackerman & Heggestad, 1997) and may also be useful in selection of graduate students. As with any selection instrument, validity, adverse impact, and job relatedness are important ethical and legal considerations.

However measured, personality and interest characteristics may predict the persistence and drive needed to complete a graduate program. Tipling (1993) examined the relationship between several noncognitive measures and graduate school performance. The correlations of positive self-concept, realistic self-appraisal, and hardiness with the criterion measure were .20, .09, and .001, respectively.

Although Tipling's (1993) correlations, as well as those from other measures, are small, it is important to remember that small validities are still useful. The argument that one should reject a predictor because the variance accounted for is only 1%, 2%, 5%, or 10% is shortsighted, not to mention potentially wrong (Ozer, 1985). Any valid predictor is superior to random selection or biased approaches. Although an overly simplistic model, the Taylor–Russell model (Taylor & Russell, 1939) illustrates the predictive power of even weak predictors. Consider the following illustration. For psychology graduate programs, Chernyeshenko and Ones (1999) found that, on average, the selection ratio is .10 (1 in 10 are admitted). For purposes of this example, assume this level of selectivity and assume that a satisfactory graduate student is one who performs above average (better than 50% of graduate students). In this scenario, given bivariate normality, even a relatively weak predictor, say one that correlates .10 with the success criterion, would increase the percentage of graduate students considered to be successful from 50% to 57%. In other words, with random selection, 50% of graduate students would score above average in terms of their performance; with the weak predictor used in student selection, 57% of graduate students would perform above average. How much improvement over the 50% graduate student success rate would there be with a predictor such as scores on the GRE Subject Tests? In this meta-analysis, the correlation between scores on the GRE Subject Tests and GGPA was .41. If the selection ratio were .10 (recall that this is the average selectivity of psychology graduate programs; Chernyeshenko & Ones, 1999), using scores on the GRE Subject Tests in selection would increase the percentage of successful graduate students from 50% to 78%. Thus, the percentage of the entering class performing satisfactorily in their course work would rise from 50% to 78%. On average, the use of this single test would produce a 28% gain in satisfactory students given a .10 selection ratio. The utility of the GRE can hardly be debated. Instead, a more fruitful debate would involve how to efficiently maximize unbiased prediction through multiple measures and information combination methods.

The burden of proof for a new predictor should lie with its proponent, who should demonstrate its incremental validity. This demonstration must take the form of multiple validations across several (large) samples and multiple criterion measures. Proof of incremental validity with multiple regression must also include appropriate corrections of R values (Campbell, 1974), because even a variable consisting of random data can create a positive ΔR (Cureton, 1950). Proof of the incremental validity of an alternative predictor would also need to address the whole battery of predictors in use, including the GRE Subject Tests and UGPA in addition to the GRE General Tests. This demonstration of a measure's validity should be especially rigorous and comprehensive when the proposal is to replace a predictor such as the GRE, which has strong validity demonstrated through massive validation efforts. To the best of our knowledge, no alternative predictor of graduate school performance has met all of these rigorous yet important requirements (with perhaps the exception of other standardized cognitive ability measures such as the Miller Analogies Test; Psychological Corporation, 1980).

Yet, even the best set of predictors is largely wasted with the commonly used clinical (subjective) data combination methods. Mechanical (algorithmic) combination of information results in superior prediction over clinical prediction (Grove & Meehl, 1996; Meehl, 1954). Despite the large body of evidence in favor of mechanical predictor combination, virtually all graduate programs rely on largely clinical combinations of quantitative and qualitative information. This approach, although superior to random selection, hamstrings the validity of admission procedures. The sizable correlations between composites of predictors presented in Table 9 occur when the data are combined mechanically via unit weighting. The research on mechanical combination of information for decision making is ubiquitous. Graduate schools should not rely on data combination methods with qualities demonstrated to be inferior more than 45 years ago.

Finally, a popular argument against mechanical combinations of data and ability testing as a whole is the notion of a nonlinear relationship between general cognitive ability and school or job performance. We are aware of no evidence to support this position including the existence of plateaus or thresholds. In the job performance domain, Coward and Sackett (1990) conducted a definitive study involving 174 independent samples with a mean sample size of 210 (a database of 36,540 individuals). They found no evidence for nonlinear relationships between ability and performance (see Jensen, 1980, and Schmidt, Ones, & Hunter 1992, for discussions of this question). Ability tests are valid predictors of performance at all levels of the trait.

In summary, the results of our investigation indicate that prior criticisms of the GRE's validity as situationally specific and useless are in error. This study examined the validity of the GRE for multiple criteria, using samples representing a wide range of academic disciplines. Our results suggest moderate correlations between GRE scores and important criterion measures, including GGPA, comprehensive examination scores, and faculty ratings of student competence. Furthermore, our results suggest that the lower correlations and much of the variability in previous research are likely to have been the result of range restriction and sampling error, respectively. The small standard deviations of the operational validities suggest it is not likely that there are variables that strongly moderate the relationships between GRE test scores and graduate school performance. Consistent with this conclusion, separate analyses of samples representing four discipline areas (humanities, social sciences, life sciences, and math–physical science), non–native English speaking students, and nontraditional students yielded results similar to those for the overall sample. Furthermore, we found no evidence to support the position that admission decisions that rely on the GRE or UGPA will result in inferior and limited graduate students. Our results indicate that the GRE is valid across disciplines for a variety of important criterion measures, and not just 1st-year GGPA, as is often assumed.

References

References marked with an asterisk indicate studies included in the meta-analysis.

Ackerman, P. L., & Heggestad, E. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin, 121,* 219–245.

*Ainslie, B. S., Anderson, L. E., Colby, B. K., Hoffman, M., Meserve, K., O'Conner, C., & Quimet, K. (1976). Predictive value of selected admission criteria for graduate nursing education. *Nursing Research, 25,* 296–299.

Alderman, D. L. (1981). *Language proficiency as a moderator variable in testing academic aptitude* (ETS Research Report RR 81-41). Princeton, NJ: Educational Testing Service.

Alderman, D. L. (1982). Language proficiency as a moderator variable in testing academic aptitude. *Journal of Educational Psychology, 74,* 580–587.

*Armstead, K. A. (1981). *An investigation of the admission requirements as predictors of success in graduate speech pathology at San Jose State University.* Unpublished doctoral dissertation, University of San Francisco, San Francisco, CA.

*Auerhahan, C. (1996). *Predictors of success in master's level nurse practitioner programs.* Unpublished doctoral dissertation, Teachers College, Columbia University, New York, NY.

*Auld, L. (1984). GRE analytical ability as an admissions factor. *Library Quarterly, 54,* 265–276.

*Baird, L. L. (1987). Do students think admissions tests are fair? Do tests affect their decisions? *Research in Higher Education, 26,* 373–388.

Barritt, L. S. (1966). Note: The consistency of first-semester college grade point average. *Journal of Educational Measurement, 3,* 261–262.

*Bean, A. (1975). The prediction of performance in an educational psychology master's degree program. *Educational and Psychological Measurement, 35,* 963–967.

Becker, B. J., & Schram, C. M. (1994). Examining explanatory models through research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 357–377). New York: Russell Sage Foundation.

Bendig, A. W. (1953). The reliability of letter grades. *Educational and Psychological Measurement, 13,* 311–321.

*Blanchard, B. E. (1977). A four-year survey of master degree graduates. *Improving College and University Teaching, 25*(2), 93–99.

*Blue, R. I., & Divilbiss, J. L. (1981). Optimizing selection of library school students. *Journal of Education for Librarianship, 21,* 301–312.

*Borg, W. R. (1963). GRE aptitude scores as predictors of GPA for graduate students in education. *Educational and Psychological Measurement, 23,* 379–382.

Borman, W. C., Hanson, M. A., Oppler, S. H., Pulakos, E. D., & White, L. A. (1993). Role of early supervisory experience in supervisor performance. *Journal of Applied Psychology, 78,* 443–449.

*Bornheimer, D. G. (1984). Predicting success in graduate school using GRE and PAEG aptitude test scores. *College and University, 60,* 54–62.

*Boudreau, R. A., Killip, S. M., MacInnis, S. H., Milloy, D. G., & Rogers, T. B. (1983). An evaluation of Graduate Record Examinations as predictors of graduate success in a Canadian context. *Canadian Psychology, 24,* 191–199.

*Bozarth, J. D., & Settles, R. B. (1980). Graduate Record Examination, race and some performance outcomes. *Rehabilitation Counseling Bulletin, 23,* 291–294.

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* (ETS Research Report 84-34). Princeton, NJ: Educational Testing Service.

Briel, J. B., O'Neill, K., & Scheuneman, J. D. (Eds.). (1993). *GRE technical manual.* Princeton, NJ: Educational Testing Service.

*Broadus, R. N., & Elmore, K. E. (1983). The comparative validities of undergraduate grade point average and of part scores on the Graduate Record Examinations in the prediction of two criterion measures in a graduate library school program. *Educational and Psychological Measurement, 43,* 543–546.

Brown, T. R., & Weaver, D. H. (1979). More than scores needed to predict graduate success. *Journalism Educator, 34,* 13–15.

*Bulechek, G. M. (1981). *Comparison of three criterion measures of success in graduate nursing education.* Unpublished doctoral dissertation, University of Iowa, Iowa City.

*Camp, J., & Clawson, T. (1979). The relationship between the Graduate Record Examination aptitude test and graduate grade point average in a master of arts in counseling program. *Educational and Psychological Measurement, 39,* 429–431.

Campbell, J. P. (1974). *A Monte Carlo approach to some problems inherent in multivariate prediction: With special reference to multiple regression* (Tech. Rep. 2002). Arlington, VA: Personnel and Training Research Programs, Office of Naval Research.

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 1, pp. 687–732). Palo Alto, CA: Consulting Psychologists Press.

Campbell, J. P., Gasser, M. B., & Oswald, F. L. (1996). The substantive nature of job performance variability. In K. R. Murphy (Ed.), *Individual differences and behavior in organizations.* San Francisco: Jossey-Bass.

Campbell, J. P., Kuncel, N. R., & Oswald, F. L. (1998, April). Predicting performance in graduate school: The criterion problem. In J. P. Campbell & D. S. Ones (Chairs), *Selection into I/O graduate programs: Focus on GRE validity.* Symposium conducted at the conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

*Capps, M. P., & DeCosta, F. A. (1957). Contributions of the Graduate Record Examinations and the National Teacher Examinations to the prediction of graduate school success. *Journal of Educational Research, 50,* 383–389.

*Case, D. O., & Richardson, J. V. (1990). Predictors of student performance with emphasis on gender and ethnic determinants. *Journal of Education for Library and Information Science, 30,* 163–182.

Chernyshenko, O. S., & Ones, D. S. (1999). How selective are psychology graduate programs? The effect of the selection ratio on GRE score validity. *Educational and Psychological Measurement, 59,* 951–961.

*Clark, H. (1968). *Graduate Record Examination correlations with grade-point averages in the Department of Education at Northern Illinois University, 1962–1966.* Unpublished master's thesis, Northern Illinois University, De Kalb.

*Clark, M. J., & Centra, J. A. (1982). *Conditions influencing the career accomplishments of Ph.D.'s* (ETS Research Report 82-18). Princeton, NJ: Educational Testing Service.

Conrad, L., Trisman, D., & Miller, R. (1977). *GRE Graduate Record*

Examinations technical manual. Princeton, NJ: Educational Testing Service.
*Conway, M. T. (1955). *The relationship of Graduate Record Examination results to achievement in the graduate school at the University of Detroit.* Unpublished master's thesis, University of Detroit.
*Cooksey, L. J. (1982). *Predictors of success in graduate school at Texas A&M University with emphasis on the analytic score on the Graduate Record Examination.* Unpublished doctoral dissertation, Texas A&M University, College Station.
*Covert, R. W., & Chansky, N. M. (1975). The moderator effect of undergraduate grade point average on the prediction of success in graduate education. *Educational and Psychological Measurement, 35,* 947–950.
Coward, W., & Sackett, P. R. (1990). Linearity of ability/performance relationships: A reconfirmation. *Journal of Applied Psychology, 75,* 297–300.
*Craven, T. F. (1981). *A comparison of admissions criteria and performance in graduate school for foreign and American students at Temple University.* Unpublished doctoral dissertation, Temple University, Philadelphia, PA.
Creager, J. A. (1966). The use of publication citations in educational research. *Journal of Educational Measurement, 3,* 243–259.
Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper & Row.
Cureton, E. E. (1950). Validity, reliability, and baloney. *Educational and Psychological Measurement, 10,* 94–96.
*Cureton, E. E., Cureton, L. W., & Bishop, R. (1949). Prediction of success in graduate study in psychology at the University of Texas. *American Psychologist, 4,* 361–362.
Dawes, R. M. (1971). A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist, 26,* 180–188.
Dawes, R. M. (1975). Graduate admissions variables and future success. *Science, 187,* 721–723.
*DeCato, C. M. (1982). Admissions criteria and later performance: Graduate Record Exam, Miller's analogies, and GPA as predictors or professional competency. *Psychological Reports, 51,* 1149–1150.
*Dole, A. A., & Baggaley, A. R. (1979). Prediction of performance in a doctoral education program by the Graduate Record Examinations and other measures. *Educational and Psychological Measurement, 39,* 421–427.
*Dollinger, S. J. (1989). Predictive validity of the Graduate Record Examination in a clinical psychology program. *Professional Psychology: Research and Practice, 20,* 56–58.
*Doris, T. J. (1990). *Prediction of performance in a doctoral education program: A replication and an extension.* Unpublished doctoral dissertation, Temple University, Philadelphia, PA.
*Duff, F. L., & Aukes, L. E. (1966). *The relationship of the Graduate Record Examination to success in the Graduate College.* Urbana: Bureau of Institutional Research and Office of Instructional Research, University of Illinois.
*Durso, R. P. (1990). *A structural analysis of factors influencing career achievement for biologists.* Unpublished doctoral dissertation, Temple University, Philadelphia, PA.
*Educational Testing Service. (1974). *The prediction of doctorate attainment in psychology, mathematics, and chemistry.* Princeton, NJ: Author.
*Educational Testing Service. (1996). *Interpreting your GRE General Test and Subject Test scores—1996–1997.* Princeton, NJ: Author.
*Educational Testing Service. (1997). *Sex, race, ethnicity, and performance on the GRE General Test: A technical report.* Princeton, NJ: Author.
Enright, M. K., & Gitomer, D. (1989). *Toward a description of successful graduate students* (GRE Board Research Report 85-17R). Princeton, NJ: Educational Testing Service.
*Enright, M. K., & Powers, D. E. (1991). *Validating the GRE analytical ability measure against faculty ratings of analytical reasoning skills* (ETS Research Report 90-22). Princeton, NJ: Educational Testing Service.
*Ewen, R. B. (1969). The GRE psychology tests as an unobtrusive measure of motivation. *Journal of Applied Psychology, 53,* 383–387.
*Federici, L., & Schuerger, J. (1974). Prediction of success in applied M. A. psychology program. *Educational and Psychological Measurement, 34,* 945–952.
*Frederiksen, N., & Ward, W. C. (1978). *Measures for the study of creativity in scientific problem solving.* Princeton, NJ: Educational Testing Service.
*Fuller, K. M. (1995). *Identification of pre-admission criteria predictive of success in graduate nurse anesthesiology programs.* Unpublished doctoral dissertation, State University of New York, Albany.
*Furst, E. J., & Roelfs, P. J. (1979). Validation of the Graduate Record Examinations and the Miller Analogies Test in a doctoral program in education. *Educational and Psychological Measurement, 39,* 147–151.
*Geisler, M. P. (1983). *The older graduate student: A descriptive study.* Unpublished doctoral dissertation, University of Wisconsin, Madison.
*Gieske, R.-M. (1990). *Academic and demographic variables related to completion status of nursing students in master's degree programs.* Unpublished doctoral dissertation, Georgia State University, Atlanta.
Glanz, J. (1996). How not to pick a physicist. *Science, 274,* 710–712.
Goldberg, E. L., & Alliger, G. M. (1992). Assessing the validity of the GRE for students in psychology: A validity generalization approach. *Educational and Psychological Measurement, 52,* 1019–1027.
*Green, S. G., & Bauer, T. N. (1995). Supervisory mentoring by advisers: Relationships with doctoral student potential, productivity, and commitment. *Personnel Psychology, 48,* 537–561.
Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law, 2,* 293–323.
*Hackman, R. J., Wiggins, N., & Bass, A. R. (1970). Prediction of long-term success in doctoral work in psychology. *Educational and Psychological Measurement, 30,* 365–374.
*Hansen, W. L. (1971). Prediction of graduate performance in economics. *Journal of Economic Education, 3,* 49–53.
*Harrison, J. D. (1989). *Predictors of success in graduate school.* Unpublished master's thesis, Texas A&I University, Kingsville.
Hartnett, R. T., & Willingham, W. W. (1980). The criterion problem: What measure of success in graduate education? *Applied Psychological Measurement, 4,* 281–291.
*Harvancik, M. J., & Golsan, G. (1986). *Graduate Record Examination scores and grade point averages: Is there a relationship?* Paper presented at the annual convention of the American Association for Counseling and Development, Los Angeles, CA. (ERIC Document Reproduction Service No. ED 270 682).
*Harvey, P. R. (1963). *Predicting graduate school performance in education.* Princeton, NJ: Educational Testing Service.
*Hatcher, L. L. (1983). *Letters of recommendation: Validity and contribution to variance in judgments about applicants.* Unpublished doctoral dissertation, Bowling Green State University, Bowling Green, OH.
Hedges, L. H., & Olkin, I. (1985). *Statistical methods for meta-analysis.* New York: Academic Press.
*Herbert, D. J., & Holmes, A. F. (1979). Graduate Record Examinations aptitude test scores as a predictor of graduate grade point average. *Educational and Psychological Measurement, 39,* 415–420.
Hirsh, H. R., Northrup, L. C., & Schmidt, F. L. (1986). Validity generalization results for law enforcement occupations. *Personnel Psychology, 39,* 399–420.
*Holmes, C. B., & Beishline, M. J. (1996). Correct classification, false positives, and false negatives in predicting completion of the Ph.D. from GRE scores. *Psychological Reports, 79,* 939–945.
Hom, P. W., Caranikas-Walker, F., Prussia, G. E., & Griffeth, R. W.

(1992). A meta-analytical structural equations analysis of a model of employee turnover. *Journal of Applied Psychology, 77,* 890–909.

*Hosford, R. E., Johnson, M. E., & Atkinson, D. R. (1984). Academic criteria, experiential background, and personal interviews as predictors of success in a counselor education program. *Counselor Education and Supervision, 23,* 268–275.

House, J. D. (1983). Effects of restriction of range on predictive validity for the Graduate Record Examination. *Psychological Reports, 53,* 710.

*House, J. D. (1989). Age bias in prediction of graduate grade point average from Graduate Record Examination scores. *Educational and Psychological Measurement, 49,* 663–666.

*House, J. D., & Johnson, J. J. (1992). Predictive validity of Graduate Record Examination scores and undergraduate grades for length of time to completion of degree. *Psychological Reports, 71,* 1019–1022.

Hoyt, D. P. (1966). College grades and adult achievement: A review of the literature. *Educational Record, 47,* 70–75.

Huitema, B. E., & Stein, C. R. (1993). Validity of the GRE without restriction of range. *Psychological Reports, 72,* 123–127.

Hunter, J. E. (1980). *Validity generalization for 12,000 jobs: An application of synthetic validity and validity generalization to the General Aptitude Test Battery (GATB).* Washington, DC: U.S. Department of Labor.

Hunter, J. E. (1983). *Test validation for 12,000 jobs: An application of job classification and validity generalization analysis to the General Aptitude Test Battery (GATB)* (Test Research Rep. No. 45). Washington, DC: U.S. Employment Service, U.S. Department of Labor.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72–98.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

*Hurdle, L. S. (1980). *An analysis of predictor variables of student achievement in general education at two regional state universities.* Unpublished doctoral dissertation, University of Connecticut, Storrs.

*Hyman, S. R. (1957). The Miller Analogies Test and the University of Pittsburgh PhD's in psychology. *American Psychologist, 12,* 35–36.

Jensen, A. R. (1980). *Bias in mental testing.* New York: Free Press.

*Kachoyeanos, M. K. (1982). *Self-esteem, internal/external control, grade point average, and Graduate Record Examination scores among graduate nursing students.* Unpublished doctoral dissertation, Northern Illinois University, De Kalb.

*Kaczmarek, M., & Franco, J. N. (1986). Sex differences in prediction of academic performance by the Graduate Record Examination. *Psychological Reports, 59,* 1159–1198.

*Kaiser, J. (1982). *The predictive validity of GRE aptitude test.* Paper presented at the annual meeting of the Rocky Mountain Research Association, Albuquerque, NM. (ERIC Document Reproduction Service No. ED 226 021)

*Kaiser, J. (1986). The validity of the GRE aptitude test for foreign students. *College Student Journal, 20,* 403–410.

*Katz, J. B. (1978). Indicators of success: Queens College Department of Library Science. *Journal of Education for Librarianship, 19,* 130–139.

*King, D. C., & Besco, R. O. (1960). The Graduate Record Examination as a selection device for graduate research fellows. *Educational and Psychological Measurement, 20,* 853–858.

*Kirnan, J. P., & Geisinger, K. F. (1981). The prediction of graduate school success in psychology. *Educational and Psychological Measurement, 41,* 815–820.

*Kluever, R. C., & Green, K. E. (1992). Prediction of achievement of doctoral students in education. *Perceptual and Motor Skills, 74,* 419–423.

*Koolboon, N. (1984). *Perceived dissertation difficulties of doctoral students in selected education departments in the United States.* Unpublished doctoral dissertation, University of Northern Colorado, Greeley.

Kuncel, N. R., Campbell, J. P., & Ones, D. S. (1998). GRE validity: Estimated or tacitly known. *American Psychologist, 53,* 567–568.

*Lamson, M. E. (1972). GRE under fire again. *Journal of Education for Librarianship, 12,* 175–178.

*Law, A. (1960). The prediction of rating of students in a doctoral training program. *Educational and Psychological Measurement, 20,* 847–851.

Law, K. S., Schmidt, F. L., & Hunter, J. E. (1994a). Non-linearity of range corrections in meta-analysis: Test of an improved procedure. *Journal of Applied Psychology, 79,* 425–438.

Law, K. S., Schmidt, F. L., & Hunter, J. E. (1994b). A test of two refinements in procedures for meta-analysis. *Journal of Applied Psychology, 79,* 978–986.

*Littlepage, G. E., Bragg, D. M., & Rust, J. O. (1978). Relations between admission criteria, academic performance and professional performance. *Teaching in Psychology, 5,* 16–20.

*Livingston, S. A., & Turner, N. J. (1982). *Effectiveness of the Graduate Record Examinations for predicting first-year grades.* Princeton, NJ: Educational Testing Service.

*Lorge, I. (1960). *Relationship between Graduate Record Examinations and Teachers College.* New York: Columbia University.

*Lust, B. L. (1981). *A study of the predictors of achievement of nurses in graduate school.* Unpublished doctoral dissertation, University of Texas at Austin.

*Madus, G. G., & Walsh, J. J. (1965). Departmental differentials in the predictive validity of the Graduate Record Examination aptitude tests. *Educational and Psychological Measurement, 25,* 1105–1110.

*Marasculio, L. A., & Gill, G. (1967). Measurable differences between successful and unsuccessful doctoral students in education. *California Journal of Educational Research, 18,* 65–70.

*Marston, A. R. (1971). It is time to reconsider the Graduate Record Examination. *American Psychologist, 26,* 653–655.

*Matthews, T. A., & Martin, D. J. (1992). Reciprocal suppression and interaction effects of age with undergraduate grades and GRE on graduate performance in a college of education. *Educational and Psychological Measurement, 28,* 453–456.

McCloy, R. A., Campbell, J. P., & Cudeck, R. (1994). A confirmatory test of a model of performance determinants. *Journal of Applied Psychology, 79,* 493–505.

Meehl, P. E. (1954). *Clinical versus statistical prediction.* Minneapolis: University of Minnesota.

*Mehrabian, A. (1969). Undergraduate ability factors in relationship to graduate performance. *Educational and Psychological Measurement, 29,* 409–419.

*Merenda, P. M., & Reilly, R. (1971). Validity of selection criteria in determining success of graduate students in psychology. *Psychological Reports, 28,* 259–266.

*Michael, J. J., Nadson, J. S., & Michael, W. B. (1983). The prediction of academic achievement in graduate study in education. *Educational and Psychological Measurement, 43,* 1133–1139.

*Michael, W. B., Jones, R. A., & Gibbons, B. D. (1960). The prediction of success in graduate work in chemistry from scores on the Graduate Record Examination. *Educational and Psychological Measurement, 20,* 859–861.

*Milner, M., McNeil, J. S., & King, S. W. (1984). The GRE: A question of validity in predicting performance in professional schools of social work. *Educational and Psychological Measurement, 44,* 945–950.

*Mohnsen, B. S. (1984). *A comparison of most and least effective graduate teaching assistants.* Unpublished doctoral dissertation, University of Southern California, Los Angeles.

Morrison, T., & Morrison, M. (1995). A meta-analytic assessment of the predictive validity of the Quantitative and Verbal components of the Graduate Record Examination with graduate grade point average representing the criterion of graduate success. *Educational and Psychological Measurement, 55,* 309–316.

*Munro, B. H. (1985). Predicting success in graduate clinical specialty programs. *Nursing Research, 34,* 54–57.

*Nagi, J. L. (1975). Predictive validity of the Graduate Record Examination and the Miller Analogies Test. *Educational and Psychological Measurement, 35,* 471–472.

*Newman, R. I. (1968). GRE scores and predictors of GPA for psychology graduate students. *Educational and Psychological Measurement, 28,* 433–436.

*Nilsson, J. E. (1995). The GRE and the GMAT: A comparison of their correlations to GGPA. *Educational and Psychological Measurement, 55,* 637–640.

Norcross, J. C., Hanych, J. M., & Terranova, R. D. (1996). Graduate study in psychology: 1992–1993. *American Psychologist, 51,* 631–643.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

*O'Connor, J., O'Connor, M., Miller, R., & Howton, B. (1984). The validity of the analytical scale in predicting first-year performance in graduate school. *Journal of Human Behavior and Learning, 1,* 9–14.

*Office of Institutional Analysis. (1966). *Correlations between admissions criteria and University of Virginia grade-point averages, Graduate School of Arts and Sciences, fall 1964.* Charlottesville: University of Virginia.

*Oldfield, K., & Hutchinson, J. R. (1997). Predictive validity of the Graduate Record Examinations with and without range restrictions. *Psychological Reports, 81,* 211–220.

*Olsen, M. (1955). *The predictive effectiveness of the Aptitude Test and the Advanced Biology Test of the GRE in the Yale School of Forestry* (Statistical Report 55-6). Princeton, NJ: Educational Testing Service.

*Omizo, M. M., & Michael, W. B. (1979). The prediction of performance in a counselor education master's degree program. *Educational and Psychological Measurement, 39,* 433–437.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology, 78,* 679–703.

Ozer, D. J. (1985). Correlation and the coefficient of determination. *Psychological Bulletin, 97,* 307–315.

*Pape, T. E. (1992). *Selected predictors of examination for professional practice in psychology scores among graduates of Western Conservative Baptist Seminary's doctoral program in clinical psychology.* Unpublished doctoral dissertation, George Fox College, Newberg, OR.

*Payne, D. A., Wells, R. A., & Clarke, R. R. (1971). Another contribution to estimating success in graduate school: A search for sex differences and comparison between three degree types. *Educational and Psychological Measurement, 31,* 497–503.

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.

Peters, L. H., Hartke, D. D., & Pohlmann, J. T. (1985). Fiedler's contingency theory of leadership: An application of the meta-analysis procedures of Schmidt and Hunter. *Psychological Bulletin, 97,* 274–285.

Premack, S. L., & Hunter, J. E. (1988). Individual unionization decisions. *Psychological Bulletin, 103,* 223–234.

Psychological Corporation. (1980). *Miller Analogies Test.* New York: Author.

Reilly, R. R. (1974). Factors in graduate student performance. *American Educational Research Journal, 13,* 125–138.

Reilly, R. R., & Warech, M. A. (1993). The validity and fairness of alternatives to cognitive tests. In L. C. Wing & B. R. Cifford (Eds.), *Policy issues in employment testing* (pp. 131–224). Boston: Kluwer.

*Rhodes, M. L., Bullough, B., & Fulton, J. (1994). The Graduate Record Examination as an admission requirement for the graduate nursing program. *Journal of Professional Nursing, 10,* 289–296.

*Roberts, P. T. (1970). *An analysis of the relationship between Graduate Record Examination scores and success in the graduate school of Wake Forest University.* Unpublished master's thesis, Wake Forest University, Winston-Salem, NC.

*Robertson, M., & Hall, E. (1964). Predicting success in graduate school. *Journal of General Psychology, 71,* 359–365.

*Robertson, S. R. C. (1993). *Admission selectors and a faculty evaluation of the personal characteristics of counseling preparation students.* Unpublished master's thesis, Northeast Missouri State University, Kirksville.

*Roscoe, J. T., & Houston, S. R. (1969). The predictive validity of GRE scores for a doctoral program in education. *Educational and Psychological Measurement, 29,* 507–509.

Roth, P. L., Be Vier, C. A., Switzer, F. S., & Schippmann, J. S. (1996). Meta-analyzing the relationship between grades and job performance. *Journal of Applied Psychology, 81,* 548–556.

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. (1993). Tacit knowledge, practical intelligence, general mental ability and job knowledge. *Current Directions in Psychological Science, 2,* 8–9.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262–274.

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

Schmidt, F. L., Hunter, J. E., & Outerbridge, A. H. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology, 71,* 432–439.

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.

Schmidt, F. L., Ones, D. S., & Hunter, J. E. (1992). Personnel selection. *Annual Review of Psychology, 43,* 627–670.

Schneider, L. M., & Briel, J. B. (1990). *Validity of the GRE: 1988–89 summary report.* Princeton, NJ: Educational Testing Service.

*Schrader, W. B. (1978). *Admissions test scores as predictors of career achievement in psychology.* Princeton, NJ: Educational Testing Service.

*Schrader, W. B. (1980). *GRE scores as predictors of career achievement in history.* Princeton, NJ: Educational Testing Service.

Shadish, W. R. (1996). Meta-analysis and the exploration of causal mediating processes: A primer of examples, methods, and issues. *Psychological Methods, 1,* 47–65.

*Sharon, A. T. (1972). English proficiency, verbal aptitude, and foreign student success in American graduate schools. *Educational and Psychological Measurement, 32,* 425–431.

*Silverston, B. E. (1984). *Classroom effectiveness and admissions criteria for intern teachers: An empirical study.* Unpublished doctoral dissertation, Claremont Graduate School, Claremont, CA.

*Sime, A. M. (1978). Prediction of success in a master's degree program in nursing. *Psychological Reports, 42,* 779–783.

*Sistrunk, F. (1961). *The GREs as predictors of graduate school success in psychology.* Unpublished manuscript.

*Sleeper, M. L. (1961). Relationship of scores on the Graduate Record Examination to grade point averages of graduate students in occupational therapy. *Educational and Psychological Measurement, 21,* 1039–1040.

*Sternberg, R. J., & Williams, W. M. (1997). Does the Graduate Record Examination predict meaningful success in the graduate training of psychologists? *American Psychologist, 52,* 630–641.

*Stricker, G., & Huber, J. T. (1967). The Graduate Record Examination and undergraduate grades as predictors of success in graduate school. *Journal of Educational Research, 60,* 466–468.

*Tanchareonrat, O. (1988). *Selected characteristics and academic achievement of international graduate students at Illinois State University.* Unpublished doctoral dissertation, Illinois State University, Normal.

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical validity of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23,* 565–578.
*Test Office, Sacramento State College. (1969). *An analysis of traditional predictor variables and various criteria of success in the master's degree programs at Sacramento State College for an experimental group who received master's degrees in the spring 1968, and a comparable control group who withdrew from their programs.* Sacramento, CA: Author.
*Thornell, J. G., & McCoy, A. (1985). The predictive validity of the Graduate Record Examination for subgroups of students in different academic disciplines. *Educational and Psychological Measurement, 45,* 415–419.
*Tipling, A. N. (1993). *Cognitive and noncognitive predictors of academic success in older women students.* Unpublished doctoral dissertation, University of Virginia, Charlottesville.
*Tully, G. E. (1962). Screening applicants for graduate study with the aptitude test of the Graduate Record Examinations. *College and University,* 51–60.
*Vacc, N. N., & Picot, R. (1984). Predicting success in doctoral study. *College Student Journal, 18*(2), 113–116.
Verhaeghen, P., & Salthouse, T. A. (1997). Meta-analysis of age-cognition relations in adulthood: Estimates of linear and nonlinear age effects and structural models. *Psychological Bulletin, 122,* 231–249.
Viswesvaran, C. (1993). *Modeling job performance: Is there a general factor?* Unpublished doctoral dissertation, University of Iowa, Iowa City.
Viswesvaran, C., & Ones, D. S. (1995). Theory testing: Combining psychometric meta-analysis and structural equations modeling. *Personnel Psychology, 48,* 865–885.
Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81,* 557–574.
*Wallace, A. D. (1952). *The predictive value of the Graduate Record Examinations at Howard University.* Unpublished master's thesis, Howard University, Washington, DC.
*Ward, W. C., & Frederiksen, N. (1977). *A study of the predictive validity of the tests of scientific thinking* (ETS Research Bulletin 77-6). Princeton, NJ: Educational Testing Service.
*Weber, J. B., & Billiland, A. R. (1942). Success in the graduate school. *Journal of Higher Education, 13,* 19–24.
*Wesche, L. E., Courtney, K., & Hauksen, C. (1984). *A study of the MAT & GRE as predictors in success in M.Ed. programs.* Seattle, WA: Pacific University. (ERIC Document Reproduction Service No. ED 310 150)
*Whetzel, D. L., & McDaniel, M. A. (1988). Reliability of validity generalization data bases. *Psychological Reports, 63,* 131–134.
*Wiggins, N., Blackburn, M., & Hackman, J. R. (1969). The prediction of first year graduate success in psychology: Peer ratings. *Journal of Educational Research, 63,* 81–85.
*Williams, J. D., Harlow, S. D., & Gab, D. (1970). A longitudinal study examining prediction of doctoral success: Grade point average as criterion, or graduation vs. nongraduation as criterion. *Journal of Educational Research, 64,* 161–164.
*Wilson, K. M. (1979). *The validation of GRE scores as predictors of first-year performance in graduate study: Report of the GRE Cooperative Validity Studies Project* (GRE Board Research Report GREB 75-8R). Princeton, NJ: Educational Testing Service. (ERIC Document Reproduction Service No. GREB-75-8R)
*Wilson, K. M. (1986). *The relationship of GRE General Test scores to first-year grades for foreign graduate students: Report of a cooperative study* (ETS Research Report 86-44). Princeton, NJ: Educational Testing Service.
*Wittmer, J., & Lister, J. L. (1971). The Graduate Record Examination, 16 PF, and counseling effectiveness. *Counselor Education and Supervision, 11,* 293.
*Yule, G., & Hoffman, P. (1990). Predicting success for international teaching assistants in a U.S. university. *TESOL Quarterly, 24,* 227–243.
*Zakzanis, K. K. (1998). The reliability of meta-analytic review. *Psychological Reports, 83,* 215–222.

Received June 30, 1998
Revision received June 22, 2000
Accepted June 22, 2000 ■