

ASSESSMENT

Standardized Tests Predict Graduate Students' Success

Nathan R. Kuncel¹ and Sarah A. Hezlett²

Accurately predicting which students are best suited for postbaccalaureate graduate school programs benefits the programs, the students, and society at large, because it allows education to be concentrated on those most likely to profit. Standardized tests are used to forecast which students will be the most successful and obtain the greatest benefit from graduate education in disciplines ranging from medicine to the humanities and from physics to law. However, controversy remains about whether such tests effectively predict performance in graduate school. Studies of standardized test scores and subsequent success in graduate school over the past 80 years have often suffered from limited sample size and present mixed conclusions of variable reliability.

Several meta-analyses have been conducted to extract more reliable conclusions about standardized tests from a variety of disciplines. To date, these review studies have been conducted on several tests commonly used in the United States: the Graduate Record Examination (GRE-T) (1), Graduate Record Examination Subject tests (GRE-S) (1), the Law School Admissions Test (LSAT) (2–4), the Pharmacy College Admissions Test (PCAT) (5), the Miller Analogies Test (MAT) (6), the Graduate Management Admissions Test (GMAT) (7), and the Medical College Admissions Test (MCAT) (8, 9).

We collected and synthesized these studies. Four consistent findings emerged: (i) Standardized tests are effective predictors of performance in graduate school. (ii) Both tests and undergraduate grades predict important academic outcomes beyond grades earned in graduate school. (iii) Standardized admissions tests predict most measures of student success better than prior college academic records do (1–5, 7, 8). (iv) The combination of tests and grades yields the most accurate predictions of success (1–4, 7, 8).

¹Department of Psychology, University of Minnesota, 75 East River Road, Minneapolis, MN 55455, USA. ²Personnel Decisions Research Institutes, 650 Third Avenue South, Suite 1350, Minneapolis, MN 55402, USA.

*Author for correspondence. E-mail: kunc001@umn.edu

Structure of Admissions Tests

Most standardized tests assess some combination of verbal, quantitative, writing, and analytical reasoning skills or discipline-specific knowledge. This is no accident, as work in all fields requires some combination of the above. The tests aim to measure the most relevant skills and knowledge for mastering a particular discipline. Although the general verbal and quantitative scales are effective predictors of student success, the strongest predictors are tests with content specifically linked to the discipline (1, 5).

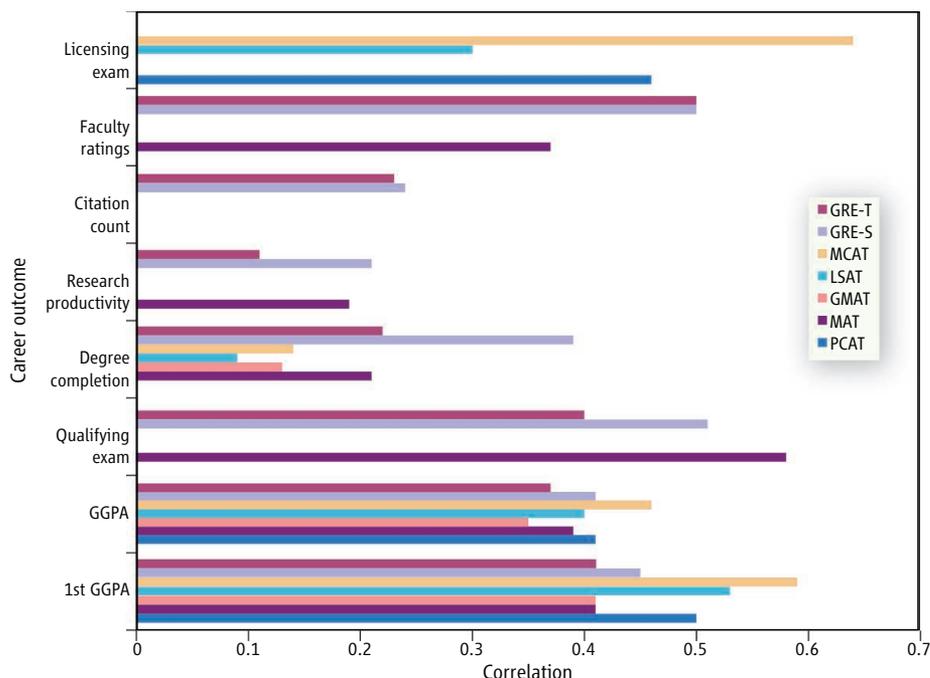
Estimating Predictive Validity

The predictive validity of tests is typically evaluated with statistics that estimate the linear relationship between predictors and a measure of academic performance. Meta-analyses synthesizing primary studies of test validity aggregate Pearson correlations. In many primary studies, the correlations are weakened by statistical artifacts, thus contributing to misinterpretation of conclusions. The first attenuating factor is the restriction of range that occurs when a

Standardized admissions tests are valid predictors of many aspects of student success across academic and applied fields.

sample is selected on the basis of a predictor variable that has a nonzero correlation with an outcome measure (10). The second attenuating factor is unreliability in the success measure resulting from inconsistency in human judgment (11). Where possible, recognized corrections were used (12) to account for these artifacts.

Research has been conducted on the correlation between test scores and various measures of student success: first-year grade point average (GPA), graduate GPA, degree attainment, qualifying or comprehensive examination scores, research productivity, research citation counts, licensing examination performance, and faculty evaluations of students. These results are based on analyses of 3 to 1231 studies across 244 to 259,640 students. The programs represented include humanities, social sciences, biological sciences, physical sciences, mathematics, and professional graduate programs in management, law, pharmacy, and medicine. For all tests across all relevant success measures, standardized test scores are positively related to subsequent meas-



Tests as predictors. Standardized test scores correlate with student success in graduate school. See table S1 for detailed data.

ures of student success [see chart, table S1, and supporting online material (SOM) text].

Utility of Standardized Tests

The actual applied utility of predictors is not easily inferred from the correlations shown in the chart. The number of correct and incorrect admissions decisions for a specific situation can be estimated from the correlation (13) (SOM text). In many cases, frequency of correct decisions can be increased from 5% to more than 30% (SOM text). In addition, correlations were converted into odds ratios using standard formulae (14) to facilitate interpretation (table S1). When an institution can be particular about whom it admits, even modest correlations can yield meaningful improvements in the performance of admitted students.

The worry that students' scores might contaminate future evaluations that, in turn, influence outcomes appears to be unfounded. The predictive validity of tests when evaluators do or do not know the individual's score is unaffected (15, 16) and, some outcomes, such as publication record, are not directly influenced by test scores.

Bias in Testing

One concern is that admissions tests might be biased against certain groups, including racial, ethnic, and gender groups. To test for bias, regression analyses of an outcome measure are compared for different groups (12, 17). If regression lines do not differ, then there is evidence that any given score on the test is associated with the same level of performance in school regardless of group membership. Overall and across tests, research has found that regression lines frequently do not differ by race or ethnic group. When they do, tests systematically favor minority groups (18–23). Tests do tend to underpredict the performance of women in college settings (24–26) (SOM text) but not in graduate school (18–20, 23).

Items from most professionally developed tests were screened for content bias and for differential item functioning (DIF). DIF examines whether performance on an item differs across racial or gender groups when overall ability is controlled. Most items do not display DIF but some content patterns have emerged over time (see SOM text). To avoid negative effects, DIF items can be rewritten, eliminated before finalizing the test, or left unscored. Research has found that the DIF effects remaining in tests have effectively zero impact on decision-making (27).

Coaching Effects in Testing

If test preparation yields large gains, then concerns about the gain's effects on differential access to higher education and the predictive validity of tests are understandable. Scores on any test can be increased to some degree because standardized tests are, like all ability tests, assessments of developed skills and knowledge. The major concern is that coaching may produce gains that are unrelated to actual skill development.

In controlled studies, gains on standardized tests used for college or graduate admissions are consistently modest. The typical magnitude for coached preparation is about 25% of one standard deviation (28–33). Longer and longer periods of study and practice are needed to attain further equal increments of improvement. Those item types that have been demonstrated to be most susceptible to coaching have been eliminated from tests (34). Test preparation or retaking does not appear to adversely affect the predictive validity of standardized tests (35–37).

Future Directions

Standardized admission tests provide useful information for predicting subsequent student performance across many disciplines. However, student motivation and interest, which are critical for sustained effort though graduate education, must be inferred from various unstandardized measures including letters of recommendation, personal statements, and interviews. Additional research is needed to develop measures that provide more reliable information about these key characteristics.

These efforts will be facilitated with more information about the actual nature of student performance. Researchers have examined a number of important outcomes but have not captured other aspects of student performance including networking, professionalism, leadership, and administrative performance. A fully specified taxonomy of student performance dimensions would be invaluable for developing and testing additional predictors of student performance.

Results from a large body of literature indicate that standardized tests are useful predictors of subsequent performance in graduate school, predict more accurately than college GPA, do not demonstrate bias, and are not damaged by test coaching. Despite differences across disciplines in grading standards, content, and pedagogy, standardized admissions tests have positive and useful relationships with subsequent student accomplishments.

References and Notes

- N. R. Kuncel, S. A. Hezlett, D. S. Ones, *Psychol. Bull.* **127**, 162 (2001).
- R. L. Linn, C. N. Hastings, *J. Educ. Meas.* **21**, 245 (1984).
- L. F. Wrightman, "Beyond FYA: Analysis of the utility of LSAT scores and UGPA for predicting academic success in law school" (RR 99-05, Law School Admission Council, Newton, PA, 2000).
- L. F. Wrightman, "LSAC national longitudinal bar passage study" (Law School Admission Council, Newton, PA, 1998).
- N. R. Kuncel *et al.*, *Am. J. Pharm. Educ.* **69**, 339 (2005).
- N. R. Kuncel, S. A. Hezlett, D. S. Ones, *J. Pers. Soc. Psychol.* **86**, 148 (2004).
- N. R. Kuncel, M. Crede, L. L. Thomas, *Acad. Manage. Learn. Educ.*, in press.
- E. R. Julian, *Acad. Med.* **80**, 910 (2005).
- R. F. Jones, S. Vanyur, *J. Med. Educ.* **59**, 527 (1984).
- R. L. Thorndike, *Personnel Selection: Test and Measurement Techniques* (Wiley, New York, 1949).
- J. E. Hunter, F. L. Schmidt, *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings* (Sage, Newbury Park, CA, 1990).
- American Educational Research Association (AERA), *Standards for Educational and Psychological Testing* (AERA, Washington, DC, 1999).
- H. C. Taylor, J. T. Russell, *J. Appl. Psych.* **23**, 565 (1939).
- J. P. T. Higgins, S. Green, Eds., *Cochrane Handbook for Systematic Review of Interventions 4.2.5* (The Cochrane Collaboration, Oxford, UK, 2005).
- B. B. Gaugler, D. B. Rosenthal, G. C. Thornton, C. Bentson, *J. Appl. Psych.* **72**, 493 (1987).
- N. Schmitt, R. Z. Gooding, R. A. Noe, M. Kirsch, *Pers. Psych.* **37**, 407 (1984).
- T. A. Cleary, *J. Educ. Meas.* **5**, 115 (1968).
- J. A. Koenig, S. G. Sireci, A. Wiley, *Acad. Med.* **73**, 1095 (1998).
- R. L. Linn, *J. Legal Educ.* **27**, 293 (1975).
- S. G. Sireci, E. Talento-Miller, *Educ. Psych. Meas.* **66**, 305 (2006).
- J. B. Vancouver, *Acad. Med.* **65**, 694 (1990).
- L. F. Wrightman, D. G. Muller, "An analysis of differential validity and differential prediction for black, Mexican American, Hispanic, and white law school students" (RR 90-03, Law School Admission Council, Newton, PA, 1990).
- H. I. Braun, D. H. Jones, "Use of empirical Bayes method in the study of the validity of academic predictors of graduate school performance" (GRE Board Professional Report 79-13P, Educational Testing Service, Princeton, NJ, 1985).
- R. Zwick, *Fair Game?* (RoutledgeFalmer, New York, 2002).
- L. J. Stricker, D. A. Rock, N. W. Burton, *J. Educ. Psych.* **85**, 710 (1993).
- J. W. Young, *J. Educ. Meas.* **28**, 37 (1991).
- S. Stark *et al.*, *J. Appl. Psych.* **89**, 497 (2004).
- B. J. Becker, *Rev. Educ. Res.* **60**, 373 (1990).
- L. F. Leary, L. E. Wrightman, "Estimating the relationship between use of test-preparation methods and scores on the Graduate Management Admission Test" (GMAC RR 83-1, Princeton, NJ, 1983).
- D. E. Powers, *J. Educ. Meas.* **22**, 121 (1985).
- S. Messick, A. Jungeblut, *Psychol. Bull.* **89**, 191 (1981).
- D. E. Powers, D. A. Rock, *J. Educ. Meas.* **36**, 93 (1999).
- J. A. Kulik *et al.*, *Psychol. Bull.* **95**, 179 (1984).
- D. E. Powers, S. S. Swinton, *J. Educ. Psych.* **76**, 266 (1984).
- M. Hojat, *J. Med. Educ.* **60**, 911 (1985).
- L. F. Wrightman, "The validity of Law School Admission Test scores for repeaters: A replication" (LSAC RR 90-02, Law School Admission Council, Newton, PA, 1990).
- A. Allalouf, G. Ben-Shakhar, *J. Educ. Meas.* **35**, 31 (1998).

Supporting Online Material

www.sciencemag.org/cgi/content/full/315/5815/1080/DC1

10.1126/science.1136618