



**ETS R&D Scientific and Policy
Contributions Series**
ETS SPC-12-02

Understanding the Impact of Special Preparation for Admissions Tests

Donald E. Powers

March 2012

Understanding the Impact of Special Preparation for Admissions Tests

Donald E. Powers
ETS, Princeton, New Jersey

ETS Research Report No. RR-12-05

March 2012

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

<http://www.ets.org/research/contact.html>

Technical Review Editor: Daniel Eignor

Technical Reviewers: Brent Bridgeman and Cathy Wendler

Copyright © 2012 by Educational Testing Service. All rights reserved.

ETS, the ETS logo, GRE, LISTENING. LEARNING. LEADING., TOEFL, and TOEIC are registered trademarks of Educational Testing Service (ETS).

SAT is a registered trademark of the College Board.



Abstract

Special preparation for tests has been a relatively contentious issue. The controversy has entailed (a) disagreement over the effectiveness of such preparation, (b) concern over unequal access to it, and (c) worries about its impact on the validity of test scores. This paper provides a brief history of ETS's involvement with, and contribution to, sorting out the issues associated with test preparation.

Key words: coaching, special test preparation, test preparation, test score validity, test familiarization, *SAT*[®] test, *GRE*[®] General Test

Foreword

Since its founding in 1947, ETS has conducted a significant and wide-ranging research program that has focused on, among other things, psychometric and statistical methodology; educational evaluation; performance assessment and scoring; large-scale assessment and evaluation; cognitive, developmental, personality, and social psychology; and education policy. This broad-based research program has helped build the science and practice of educational measurement, as well as inform policy debates.

In 2010, we began to synthesize these scientific and policy contributions, with the intention to release a series of reports sequentially over the course of 2011 and 2012. These reports constitute the *ETS R&D Scientific and Policy Contributions Series*.

The inaugural report in the series was published in spring 2011 and was a re-issue of a paper by Samuel Ball that documented the vigorous program of evaluation research conducted at ETS in the 1960s and 1970, which helped lay the foundation for this fledgling field. The second report in the series, by Donald Rock, appeared in early 2012 and reviewed ETS's contribution to several large-scale longitudinal assessments over the years, ranging from the National Longitudinal Study of the High School Class of 1972 (NLS-72) to the Early Childhood Longitudinal Studies (ECLS).

The latest report in the series, by Donald Powers, examines the role of special test preparation. Test preparation is a matter of some controversy, raising such issues as effectiveness of coaching programs and products, unequal access to them, and concerns about the impact of preparation on the validity of test scores. Powers documents ETS's contributions in addressing these issues, including analysis of key features of test preparation, research into the effects of various types of preparation, and creation of tests that yield meaningful scores in the face of legitimate as well as questionable attempts to improve test performance.

One of the most important contributions in this area is the methodological rigor ETS researchers have brought to the field, demonstrating the feasibility of conducting experimental studies of the effects of test preparation. This rigor has resulted in the introduction of more sophisticated methods for controlling self-selection bias in nonexperimental studies of the effects of coaching. Powers shows that ETS research on test preparation has been more than an academic exercise. It has resulted in significant—even dramatic—modifications to several ETS

tests. Considerations about test preparation now factor into the *design* of ETS tests, well before they are ever administered to test takers.

Donald Powers is currently a managing principal research scientist in the Research & Development Division at ETS. Much of his research during his long career at ETS has concentrated on understanding the sources of variation in performance on national admissions tests through studies of the effects of test familiarization and special test preparation or coaching, especially as they affect test interpretation and use. He has also served as research coordinator for a number of ETS programs, including the *TOEFL*[®], *GRE*[®], and *TOEIC*[®] tests.

Future reports in the *ETS R&D Scientific and Policy Contributions Series* will focus on other major areas of research and education policy in which ETS has played a role.

Ida Lawrence
Senior Vice-President
Research & Development Division
ETS

Overview

By examining unique developments and singular advancements, it is possible to sort the history of educational and psychological testing into a number of distinct phases. One topic that seems to permeate all stages, however, is the question of how best to *prepare* for such tests. This paper documents some of ETS's contributions to understanding the role of test preparation in the testing process. These contributions include (a) analyzing key features of test preparation, (b) understanding the effects of various sorts of preparation on test performance, and (c) devising tests that will yield meaningful scores in the face of both legitimate as well as questionable attempts to improve test-taker performance. The paper begins with a definition of special test preparation and then elaborates on its significance. Next, it examines the nature of interest in the topic. Finally, it explores ETS Research and Development (R&D) contributions to explicating the issues associated with special test preparation.

Definitions

The first issue that one encounters when discussing test preparation is terminology. This terminology applies both to the tests that are involved and to the kinds of preparation that are directed at test takers. Most of the research described below pertains to several tests that are designed to measure academic abilities (e.g., verbal and quantitative reasoning abilities) that develop relatively slowly over a significant period of time. This improvement occurs as a result of both formal schooling as well as other less formal experiences outside of school. Thus, to varying degrees, all students who take these kinds of tests receive highly relevant (but certainly differentially effective) preparation that should improve the skills and abilities being tested.

With respect to preparation, we have chosen here to use the word *special* to refer to a particular category of test preparation that focuses on readying test takers for a specific test. This special preparation may be of different sorts. For example, test familiarization is designed to ensure that prospective test takers are well versed in the general skills required for test taking and to help them gain familiarity with the procedures that are required to take a particular test. This type of preparation may entail, for instance, exposing test takers to the kinds of item formats they will encounter, making certain that they know when to guess, and helping them learn to apportion their time appropriately. Special preparation of this sort is generally regarded as desirable, as it presumably enables individuals to master the mechanics of test taking, thereby

freeing them to focus on, and accurately demonstrate, the skills and abilities that are being assessed.

Coaching, on the other hand, has had a decidedly more negative connotation insofar as it is typically associated with short-term efforts aimed at teaching test-taking strategies or “tricks” to enable test takers to “beat the test;” that is, to take advantage of flaws in the test or in the testing system (e.g., never choose a particular answer choice if a question has these characteristics...). As Messick (1982) has noted, however, the term *coaching* has often been used in a variety of ways. At one extreme, it may signify short-term cramming and practice on sample item types, while on the other it may denote long-term instruction designed to develop the skills and abilities that are being tested. In practice, the distinctions among (a) relevant instruction, (b) test familiarization, and (c) coaching are sometimes fuzzy, as many programs contain elements of each type of preparation.

Significance of Special Test Preparation

Messick (1982) noted three ways in which special preparation may improve test scores. Each of these ways has a very different implication for score use. First, like real instruction, some types of special test preparation may genuinely improve the skills and abilities being tested, thereby resulting in higher test scores also. This outcome should have no detrimental effect on the validity of scores.

Second, some special test preparation (or familiarization) may enhance general test-taking skills and reduce test anxiety, thereby increasing test scores that may otherwise have been inaccurately low indicators of test takers’ true abilities. Insofar as this kind of preparation reduces or eliminates unwanted sources of test difficulty, it should serve only to improve score validity.

The third possibility is that if it entails the teaching of test-taking tricks or other such strategies, special test preparation may increase test scores without necessarily improving the underlying abilities that are being assessed. A likely result is inaccurately high test scores and diminished score validity.

Finally, along with score validity, equity is often at issue in special test preparation, as typically not all students have equal opportunity to benefit in the ways described above. If special preparation is effective, its benefits may accrue only to those who can afford it.

Interest in Special Test Preparation

At first blush, the issue of special test preparation might seem to be of interest mainly to a relatively small group of test developers and psychometricians. Historically, however, attention to this topic has been considerably more widespread. Naturally, test takers (and for some tests, their parents) are concerned with ensuring that they are well prepared to take any tests that have high-stakes consequences. However, other identifiable groups have also shown considerable interest in the topic.

For instance, concern is clearly evident in the professional community. The current version of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999) suggests a need to establish the degree to which a test is susceptible to improvement from special test preparation (Standard 1.9: “If a test is claimed to be essentially unaffected by practice and coaching, the sensitivity of test performance to change with these forms of instruction should be documented,” p. 19). In addition, a previous edition of *Educational Measurement* (Linn, 1989), perhaps the most authoritative work on educational testing, devoted an entire chapter to special test preparation (Bond, 1989).

General public interest is apparent also, as coaching has been the subject of numerous articles in the popular media (e.g., “ETS and the Coaching Cover Up,” Levy, 1979). One study of the effects of coaching (Powers & Rock, 1999) was even a topic of discussion on a prominent national television show when the host of the *Today Show*, Matt Lauer, interviewed College Board Vice President Wayne Camara.

Besides being of general interest to the public, ETS coaching studies have also had a major impact on testing policy and practice. For example, in the early 1980s a previously offered section of the *GRE*[®] General Test (the analytical ability measure) was changed radically on the basis of the results of a GRE Board-sponsored test preparation study (Powers & Swinton, 1984).

As a final indication of the widespread interest in the topic, in the late 1970s the U.S. Federal Trade Commission (FTC) became so troubled by the possibly misleading advertising of commercial coaching companies that it launched a major national investigation of the efficacy of such programs (Federal Trade Commission, 1978, 1979). As described below, ETS contributed in several ways to this effort.

Studying the Effects of Special Test Preparation

What follows is an account of several key ETS contributions to understanding the role and effects of special test preparation. The account is organized within each of the two major testing programs on which special test preparation research has concentrated, the *SAT*[®] test and the *GRE* General Test.

The SAT

The College Board position. The *effectiveness* of special test preparation has long been a contentious issue. Perhaps a reasonable place to begin the discussion is with the publication of the College Board's stance on coaching, as proclaimed by Board's trustees in "Effects of Coaching on Scholastic Aptitude Test Scores" (College Entrance Examination Board, 1965). This booklet summarized the (then) relatively few studies of coaching for the SAT and concluded that intensive coaching was, at best, likely to yield "negligible" results (p. 8). At the time, there was, at least in some circles, considerable skepticism about the Board's position.

Early studies. The first serious disagreement with the Board's stance seems to have come with the completion of a study by ETS researchers Evans and Pike (1973), who demonstrated that two SAT quantitative item types being considered for inclusion in the SAT were susceptible to improvement through special preparation—in particular, to the Saturday morning test preparation classes that the researchers designed for implementation over a 7-week period. The researchers' best estimate of effects was about 25 points on the 200–800 SAT Math (SAT-M) scale.

Besides the significant program of instruction that Evans and Pike developed, another particularly noteworthy aspect of this effort was the researchers' ability to implement a true experimental design. Students were randomly assigned to either (a) one of three treatment groups, each of which focused specifically on a different item type, or (b) a comparison condition that involved only more general test-taking skills. Previously, virtually no such studies had successfully carried out a true experiment.

At least partly because of the Evans and Pike (1973) study, interest also increased in the effects of special preparation for the verbal section of the SAT. The College Board subsequently funded ETS researchers to study the effectiveness of special secondary school programs geared to improving SAT Verbal (SAT-V) scores (Alderman & Powers, 1980). A contribution here was that instead of relying on strictly observational methods or quasi-experimental designs, the

investigators were able, through careful collaboration with a set of secondary schools, to exert a reasonably strong degree of experimental control over *existing* special preparation programs, assigning students randomly to treatment or control groups. This task was accomplished, for example, by taking advantage of demand for preparation that, in some cases, exceeded the schools' ability to offer it. In other cases, it was possible to simply delay preparation for randomly selected students. The results suggested that secondary school programs can affect SAT-V scores, albeit modestly, increasing them by about 4 to 16 points on the 200–800 SAT-V scale.

Test familiarization. About the same time, the College Board, realizing the need to ensure that all test takers were familiar with the SAT, developed a much more extensive information bulletin than had been available previously. The new booklet, called *Taking the SAT*, contained extensive information about the test and about test-taking strategies, a review of math concepts, and a full-length practice SAT. Much to its credit, the Board was interested not only in offering the more extensive preparation material, but also in learning about its impact, and so it commissioned a study to assess the booklet's effects on both test-taking behavior and test scores (Powers & Alderman, 1983). The study was a true experiment in which a randomly selected group of SAT registrants received a prepublication version of the new booklet. Subsequently, their test performance was compared with that of an equivalent randomly selected group of test takers who had not received the booklet. (Only high school juniors were included in the study, partly to ensure that, should the booklet prove effective in increasing scores, all students in the cohort would have the opportunity to benefit from it before they graduated.)

The results showed increases in knowledge of appropriate test-taking behavior (e.g., when to guess), decreased anxiety, and increased confidence. There were no statistically significant effects on SAT-V scores but a small, significant effect on SAT-M scores of about 8 points.

Federal interest. Perhaps the single most significant factor in the rising interest in coaching and test preparation was the involvement of the U.S. Federal Trade Commission (FTC). The FTC became increasingly interested in the veracity of claims being made by commercial coaching companies, which promised to increase SAT takers' scores by hundreds of points. The issue became so important that the FTC eventually undertook its own study to investigate the effectiveness of commercial coaching programs.

Ultimately, both ETS and several of the major commercial coaching companies cooperated with the FTC investigation. ETS provided students' SAT scores, and the coaching companies provided information about students' enrollment in their programs. FTC researchers analyzed the data and eventually issued a report, finding the effects of commercial coaching for the SAT to be statistically significant—in the range of 20–30 points for both SAT-V and SAT-M at the most effective of the coaching schools that were studied (Federal Trade Commission, 1978, 1979; Sesnowitz, Bernhardt, & Knain, 1982). Needless to say, the study attracted considerable attention.

ETS responded to the FTC's findings as follows. Sam Messick, then Vice President for Research at ETS, assembled a team of researchers to take a critical look at the methods the FTC had used and the conclusions it had reached. Messick and his team critiqued the FTC's methodology and, in order to address some serious flaws in the FTC analyses, reanalyzed the data that had been collected. Various analyses were employed to correct mainly for test taker self-selection in attending coaching programs.

Messick's contribution was released as a monograph titled, "The Effectiveness of Coaching for the SAT: Review and Reanalysis of Research from the Fifties to the FTC" (Messick, 1980). In the book, Messick summarized and critiqued previous research on coaching, and several ETS researchers offered their critiques of the FTC study. Most importantly, the researchers conducted several reanalyses of the data obtained from the FTC. ETS consultant Thomas Stroud reanalyzed the data, controlling for a variety of background variables and found results similar to those reported by the FTC. By considering *PSAT* scores, as well pre- and postcoaching SAT scores, ETS researcher Don Rock was able to apply a differential growth model to the FTC data. His analysis showed that, at least for SAT-V scores, some of the difference between the posttest SAT scores of coached and uncoached test takers could be attributed simply to the faster growth exhibited by coached students rather than to any specific effect of coaching. The results of the various ETS analyses differed somewhat, but in total they revealed that only one of the three coaching schools had a significant impact on SAT scores—about 12–18 points on the SAT-V scale and about 20–30 points on the SAT-M scale.

One of the main lessons from the critique and reanalysis of the FTC study was stated by Messick (1980) in the preface to the report. Messick wrote that the issue of the effectiveness of coaching for the SAT is much more complicated than the simplistic question of whether

coaching works or not. Coaching in and of itself is not automatically to be either rejected or encouraged. Rather, it matters what materials and practices are involved, at what cost in student time and resources, and with what effect on student skills, attitudes, and test scores. (p. v)

Messick's (1980) insight was that complex issues, like the coaching controversy, are rarely ever usefully framed as simple either/or, yes/no questions. Rather, those questions turn out to involve degrees and multiple factors that need to be appreciated and sorted out. As a consequence, the answer to most questions is usually not a simple "yes" or "no," but more often a sometimes frustrating, "it depends." The task of researchers, then, is usually to determine, as best they can, the factors on which the effects depend.

Extending lessons learned. Messick followed through with this theme by analyzing the relationship of test preparation effects to the duration or length of test preparation programs. He published these results in the form of a meta-analysis (Messick & Jungeblut, 1981), in which the authors noted "definite regularities" (p. 191) between SAT coaching effects and the amount of student contact time in coaching programs. On this basis, Messick and Jungeblut concluded that the size of the effects being claimed by coaching companies could probably be obtained only with programs that were tantamount to full-time schooling.

Powers (1986) followed Messick and Jungeblut's (1981) lead by reviewing a variety of other features of test preparation and coaching programs, and relating these features to the size of coaching effects. The advance here was that instead of focusing on the features of coaching programs, Powers analyzed the characteristics of the *item types* that comprised a variety of tests—for instance, how complex their directions were, whether they were administered under timed or untimed conditions, and what kinds of formats they employed. The results suggested that some features of test items (e.g., the complexity of directions) did render them more susceptible to improvement through coaching and practice than did others.

Several of the studies that Powers reviewed were so-called within-test practice studies, which were conducted by ETS statistical analysts (e.g., Faggen & McPeck, 1981; Swinton, Wild, & Wallmark, 1983; Wightman, 1981). This innovative method involved trying out new test item types in early and later sections of the same test form. Then, differences in performance were compared for these early and later administered items. For some item types, it was routinely noticed that examinees performed better on new items that appeared later in the test, after earlier appearances of items of that type. A large within-test practice effect was viewed as a sufficient

condition to disqualify a proposed new item type from eventual operational use. The rationale was the following: If an item type exhibited susceptibility to simple practice *within* a single test session, surely it would be *at least* as susceptible to more intensive coaching efforts.

Studying the revised SAT. In 1994, a revision of the SAT was introduced. Many of the changes suggested that the revision should be even less susceptible to coaching than the earlier version. However, claims being made by coaching companies did not subside. For example, the January, 8, 1995, issue of the Philadelphia *Inquirer* proclaimed “New SAT proves more coachable than old.” At least partly in response to such announcements, the College Board sponsored research to examine the effects of commercial coaching on SAT scores. Powers and Rock (1999) surveyed SAT takers about their test preparation activities, identifying a subset of test takers who had attended commercial coaching programs. Although the study was observational in nature, the researchers obtained a wide variety of background information on test takers and used this information to control statistically for self-selection effects. This approach was necessary, as it was widely acknowledged that coached and uncoached students differ on numerous factors that are also related to SAT scores. One of the differences noted by Powers and Rock, and controlled in their analysis, was that coached test takers were more likely than their uncoached counterparts to have engaged in a variety of *other* test preparation activities (e.g., self-study of various sorts), which may also have affected SAT scores. Several alternative analyses were employed to control for self-selection effects, and although each of the analyses produced slightly different estimates, all of them suggested that the effects of coaching were far less than was being alleged by coaching enterprises—perhaps only a quarter as large as claimed.

The alternative analyses yielded coaching effect estimates of 6–12 for SAT-V and 13–26 points for SAT-M. When analyses were undertaken separately for major coaching companies, the results revealed SAT-V effects of 12–19 points for one company and 5–14 points for another. The effects for SAT-M were 5–17 and 31–38, respectively, suggesting that the two programs were differentially effective for the two portions of the SAT.

The results of the study were featured in a New York *Times* article (Bronner, 1998). The article quoted Professor Betsy Jane Becker, who had reviewed numerous SAT coaching studies (Becker, 1990), as saying that the study was “perhaps the finest piece of coaching research yet published” (p. A23). This assessment may of course reflect either a regard for the high quality of the study or, on the other hand, concern about the limitations of previous ones.

The GRE General Test

Although the SAT program has been a major focus of test preparation and coaching studies, the GRE Board has also sponsored a number of significant efforts by ETS researchers. For instance, the GRE program revised its General Test in the late 1970s, introducing an analytical ability measure to complement the long-offered verbal and quantitative reasoning measures (Powers & Swinton, 1981). Concurrently, the GRE Board sponsored several studies to examine the susceptibility of the new measure to coaching and other forms of special test preparation. Swinton and Powers (1983) designed a brief course to prepare students for the new analytical section of the GRE General Test and offered it to a small group of volunteer GRE test takers at a local university. Controlling for important pre-existing differences between groups, they compared the postcourse GRE performance of these specially prepared individuals with that of all other GRE test takers at the same university. They found that the specially prepared group did much better on the analytical section (by about 66 points on the 200–800 scale) than did the larger comparison group, even after controlling for differences in the GRE verbal and quantitative scores of the two groups.

Powers and Swinton (1984) subsequently packaged the course and used it in a true experimental study in which a randomly selected sample of GRE test takers received the course materials by mail. A comparison of the test scores of the prepared sample with those of a randomly selected equivalent sample of nonprepared GRE test takers revealed score improvements that were nearly as large (about 53 points with about 4 hours of self-preparation) as those observed in the face-to-face classroom preparation. A major implication of this latter study was that test preparation designed for self-study by test takers themselves was a viable alternative to more expensive, formal face-to-face interventions. The ramifications for fairness and equity were obvious. However, although the researchers were relatively sanguine about the prospects for ensuring that all examinees could be well prepared for the “coachable” item types on the GRE, the GRE Board took a conservative stance, deciding instead to remove the two most susceptible item types from the analytical ability measure.

Data collected in the studies of the GRE analytical measure were also used to gauge the effectiveness of formal commercial coaching for the verbal and quantitative sections (Powers, 1985a). That is, since the analytical measure had been shown to be coachable, it could serve as a baseline against which to judge the coachability of the other test sections.

For this analysis, Powers identified test takers who had attended formal coaching programs for any or all of the GRE test sections. For the analytical ability section, the analysis revealed a strong relationship between the effect of coaching and its duration (in terms of hours devoted to instruction). However, applying the same methodology to the verbal and quantitative sections revealed little if any such relationship, contrary to claims being made by commercial coaching firms. Increasing the duration of preparation for the verbal and quantitative GRE measures did not produce commensurate increases in scores for these two measures.

Effects on relationships of test scores with other measures. While Messick (1982) provided an insightful *logical* analysis of the ways in which special test preparation may impact validity, there appears to have been little *empirical* research to demonstrate how such practices may affect, for example, the relationship of test scores to other relevant measures. An exception is a study by Powers (1985b), who examined the relationship of GRE analytical ability scores, obtained under 10 different randomly assigned test preparation conditions, to indicators of academic performance. Each of the various test preparation conditions was designed, mainly, to help test takers become familiar with each of several novel analytical ability item types. The results suggested that the more time test takers devoted to using the test preparation materials, the stronger the relationship was between academic performance and scores on the GRE analytical ability measure. Specifically, over the ten treatment groups, the correlation between (a) GRE analytical ability score and (b) undergraduate grade point average in the final two years of undergraduate study increased according to mean time devoted to preparing for the analytical measure ($r = .70, p < .05$). In addition, correlations of GRE analytical ability scores with GRE verbal and quantitative scores decreased, though not significantly, with increasing amounts of test preparation. Thus, both the convergent and (possibly) the discriminant aspects of construct validity of test scores may have been enhanced.

Summary

ETS R&D has made several contributions to understanding the effects of special test preparation and coaching on (a) test-taking behavior, (b) test performance, and (c) test validity. First, ETS researchers have brought more methodological rigor to the field by demonstrating the feasibility of conducting experimental studies of the effects of test preparation. Rigor has also been increased by introducing more sophisticated methods for controlling self-selection bias in nonexperimental studies.

Moreover, ETS researchers have evaluated the effects of a variety of different types of test preparation: formal commercial coaching, school-offered test preparation programs, and test sponsor-provided test familiarization. With respect to the last type, a significant portion of the ETS-conducted research has focused on making certain that *all* test takers are well prepared, not just those who can afford extensive coaching. Along these lines, researchers have evaluated the effects of test familiarization and other means of test preparation that can be offered, usually remotely for independent study, to *all* test takers. Both secondary and postsecondary student populations have been studied.

Thanks largely to Messick (1980, 1981, 1982), the question of the effectiveness of coaching and test preparation has been reformulated—that is, extended beyond the search for a simple dichotomous yes/no answer to the oversimplified question “Does coaching work?” Partly as a result, researchers now seem more inclined to examine the components of test preparation programs in order to ascertain the particular features that are implicated in its effectiveness.

ETS researchers have also stressed that every test is typically composed of a variety of item types and that some of these item types may be more susceptible to coaching and practice than others. In this vein, they have determined some of the features of test item types that seem to render them more or less susceptible. As a consequence, there is now a greater realization that it is insufficient to simply consider the coachability of a test as a whole, but rather it is necessary to consider the characteristics of the various item types that comprise it.

In addition, at least in the view of the serious scientific community, if not among the general public, a more accurate estimate of the true value of commercial coaching programs now exists. Consumers have information to make more informed choices about whether to seek commercial coaching, for instance. The true effect of coaching on test performance seems neither as negligible as some have claimed nor as large as has been advertised by the purveyors of coaching services.

Most of the studies of coaching and test preparation have focused on the extent to which these practices cause spurious test score improvement. However, although relatively rare, ETS researchers have also examined, in both a logical and an empirical manner, the effects of test preparation and coaching on the empirical relationships of test scores to other indicators of developed ability.

Finally, ETS research on test preparation has been more than an academic exercise. It has resulted in significant—even dramatic—modifications to several tests that ETS offers. These changes are perhaps the clearest example of the impact of ETS’s research on test preparation. However, there have, arguably, been more subtle effects as well. Now, when new assessments are being developed, the potential coachability of proposed new test item types is likely to be a factor in decisions about the final composition of a test. Considerations about test preparation figure into the *design* of tests, well before these tests are ever administered to test takers.

References

- Alderman, D. L., & Powers, D. E. (1980). The effects of special preparation on SAT verbal scores. *American Educational Research Journal*, 17, 239–251.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Becker, B. J. (1990). Coaching for the Scholastic Aptitude Test: Further synthesis and appraisal. *Review of Educational Research*, 60, 373–417.
- Bond, L. (1989). The effects of special preparation on measures of scholastic ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 429–444). New York, NY: Macmillan,.
- Bronner, E. (1998, November 24). Study casts doubt on the benefits of S.A.T.-coaching courses. *The New York Times National*, p. A23.
- College Entrance Examination Board. (1965). *Effects of coaching on Scholastic Aptitude Test scores*. New York, NY: Author.
- Evans, F. R., & Pike, L. W. (1973). The effects of instruction for three mathematics item formats. *Journal of Educational Measurement*, 10, 257–272.
- Faggen, J., & McPeck, M. (1981, April). *Practice effects for four different item types*. Paper presented at the annual meeting of the National Council on Measurement in Education, Los Angeles, CA.
- Federal Trade Commission, Boston Regional Office. (1978, September). *Staff memorandum of the Boston Regional Office of the Federal Trade Commission: The effects of coaching on standardized admission examinations*. Boston, MA: Author.
- Federal Trade Commission, Bureau of Consumer Protection. (1979). *Effects of coaching on standardized admission examinations: Revised statistical analyses of data gathered by Boston Regional Office, Federal Trade Commission*. Washington, DC: Author.
- Levy, S. (1979, March). ETS and the coaching cover-up. *New Jersey Monthly*, 3(4), 50–54, 82–89.
- Linn, R. L. (Ed.). (1989). *Educational measurement* (3rd ed.). New York, NY: Macmillan.

- Messick, S. (1980). *The effectiveness of coaching for the SAT: Review and reanalysis of research from the fifties to the FTC*. Princeton, NJ: ETS.
- Messick, S. (1981). The controversy over coaching: Issues of effectiveness and equity. In B. F. Green (Ed.), *Issues in testing: Coaching, disclosure, and ethnic bias*. San Francisco, CA: Jossey-Bass.
- Messick, S. (1982). Issues of effectiveness and equity in the coaching controversy: Implications for educational testing and practice. *Educational Psychologist, 17*, 67–91.
- Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin, 89*, 191–216.
- Powers, D. E. (1985a). Effects of coaching on GRE Aptitude Test scores. *Journal of Educational Measurement, 22*, 121–136.
- Powers, D. E. (1985b). Effects of test preparation on the validity of a graduate admissions test. *Applied Psychological Measurement, 9*, 179–190.
- Powers, D. E. (1986). Relations of test item characteristics to test preparation/test practice effects: A quantitative summary. *Psychological Bulletin, 100*, 67–77.
- Powers, D. E., & Alderman, D. L. (1983). Effects of test familiarization on SAT performance. *Journal of Educational Measurement, 20*, 71–79.
- Powers, D. E., & Rock, D. A. (1999). Effects of coaching on SAT I: Reasoning scores. *Journal of Educational Measurement, 36*, 93–118.
- Powers, D. E., & Swinton, S. S. (1981). Extending the measurement of graduate admission abilities beyond the verbal and quantitative domains. *Applied Psychological Measurement, 5*, 141–158.
- Powers, D. E., & Swinton, S. S. (1984). Effects of self-study for coachable test item types. *Journal of Educational Psychology, 76*(2), 266–278.
- Sesnowitz, M., Bernhardt, K. L., & Knain, D. M. (1982). An analysis of the impact of commercial test preparation courses on SAT scores. *American Educational Research Journal, 19*, 429–441.
- Swinton, S. S., & Powers, D. E. (1983). A study of the effects of special preparation on GRE analytical scores and item types. *Journal of Educational Psychology, 75*, 104–115.

Swinton, S. S., Wild, C. L., & Wallmark, M. M. (1983). *Investigating practice effects on item types in the Graduate Record Examinations aptitude test* (ETS Research Report No. RR-82-56). Princeton, NJ: ETS.

Wightman, L. E. (1981, April). *GMAT within-test practice effects studies*. Paper presented at the annual meeting of the National Council of Measurement in Education, Los Angeles, CA.