Review of Educational Research
Month 202X, Vol. XX, No. X, pp. 1–45
DOI: 10.3102/00346543251360775

The Impact of Test Preparation on Performance of Large-Scale Educational Tests: A Meta-analysis of Experimental Studies

Zhanxin Hao
Tsinghua University

Jo-Anne Baird
University of Oxford

Yasmine EI Masri
Ofqual

Kit Double
The University of Sydney

Background: *Test preparation has been widely used to optimize test-takers' potential on large-scale educational tests; meanwhile, people questioned its effect and raised concerns that it may narrow students' learning.*

Methods: *This meta-analysis provided a statistical synthesis of existing experimental and quasi-experimental studies on the effects of test preparation interventions on students' test performance on large-scale educational tests.*

Results: *Results from 28 included studies suggested that students' test performance can be significantly improved by test preparation ($g = .26$, 95% CI = .10–.42). Using workbooks, developing students' socio-affective strategies, and teaching test-taking skills were found to be significant moderators of the test preparation effect. Little evidence of practice effect (i.e., using sample items and practice tests) on test preparation was found in the current study.*

Discussion: *The effect of test preparation could be attributed to an increase in domain-specific knowledge or test-specific cognitive skills. However, the current framework of test preparation effect was mainly built upon evidence of the retest effect on cognitive ability tests. We believe that learning and cognition theories provide hints as to how to explain findings in this study, though more evidence from future research is needed.*

Keywords: achievement, learning processes/strategies, school/teacher effectiveness, testing, test theory/development, tutoring, experimental research, experimental design, meta-analysis, quasi-experimental analysis, test preparation effect, test performance, large-scale educational tests, meta-analysis

Introduction

Large-scale educational tests, such as college admission tests, are an inherent part of education systems throughout the world. Test scores have been used tremendously by educational authorities internationally and nationally for making decisions related to gatekeeping, accountability, and policymaking. The increasing power of test scores has spawned an ever-growing demand for test preparation inside or outside schools aimed at helping students to prepare for tests (Bray, 1999; Buchmann et al., 2010; Ross, 2008). Some schools provide students with test preparation courses to achieve greater accountability. This approach ensures that the schools are directly responsible for equipping students with the knowledge and skills needed to excel in exams, reflecting the schools' commitment to students' academic excellence. Besides, a large number of commercial educational institutions have built well-established test preparation training systems that not only provide group courses and private tutorials but also publish practice books or even develop online training systems. The underlying rationale for these test preparation activities is that test preparation is conducive to maximizing test results.

Although the test preparation industry is flourishing, there has been insufficient research exploring how, and to what extent, test preparation is effective. Existing literature seems to converge on the notion that test preparation, in most cases, has a positive effect on test performance (e.g., Appelrouth et al., 2017; Briggs, 2001, 2009; Bunting & Mooney, 2001; Hausknecht et al., 2007; Powers & Rock, 1999). However, the effect shown in those studies is inconsistent and inconclusive. Some studies detected a moderate to strong effect size (e.g., Hausknecht et al., 2007, $d = .70$), whereas others showed weaker or even negative effects (e.g., Briggs, 2009; McGaghie et al., 2004). In this case, an overarching quantitative synthesis of existing empirical research regarding the effect of test preparation, which would make it possible to draw a general conclusion, is needed. Therefore, we present a meta-analysis to investigate the effect of test preparation activities in educational settings on test performance of large-scale tests, in comparison with no preparation or "business-as-usual" conditions. This study also provides valuable insights into possible differential effects of test preparation by including specific characteristics of test preparation intervention design, strategy and material use, and the nature of target test in a moderator analysis.

Evaluating the Evidence of Test Preparation

Working Definition

Various terms have been used in the literature to describe teachers' and students' test preparation behaviors, such as coaching, teaching to the test, test-oriented instruction, test familiarization, private tutoring, test-wiseness, etc. In this

study, we use "test preparation" as an inclusive term, which refers to "any intervention procedure specifically undertaken to improve test scores, whether by improving the skills measured by the test or by improving the skills for taking the test, or both" (Messick, 1982, p. 70). In this broad definition, all learning activities, no matter whether they were driven by teachers or students, as long as the main intention was to improve performance in a specific test, could be considered as test preparation activities. Based on this understanding, the indicator of the effects of test preparation activities could be either the observed gain in a test-taker's test scores after some form of test preparation has taken place, or the gap of test performance between those who engaged in test preparation with their unprepared cohorts (Anastasi, 1981; Briggs, 2001; Messick, 1982).

A large number of empirical studies have reported teachers and students' engagement in test preparation activities (e.g., Anastasi, 1981; Appelrouth et al., 2017; Briggs, 2009; Lai & Waltman, 2008). In this article, we categorize test preparation activities as either student-driven or instructor-driven based on Briggs's (2009) classification. Student-driven test preparation has various forms, including but not limited to self-studying test preparation booklets, using online materials or computer programs, practicing sample test items or taking practice tests, and asking about the experience from previous test-takers. Student-driven test preparation has been consistently reported as the most used test preparation method (Briggs, 2009). In comparison, instructor-driven test preparation mainly comprises of coaching or training courses on admission tests or other high-stakes tests in both formal schools and commercial educational institutes (Bangert-Drowns, Kulik, & Kulik, 1983; Briggs, 2009).

Previous Empirical Research on the Test Preparation Effect

Although the effect of test preparation has been investigated by a considerable number of empirical studies, only a few used a (quasi) experimental design. Most studies used self-report surveys to investigate participants' test preparation histories, then compared test performance of prepared students to that of unprepared students (e.g., Briggs, 2001; Domingue & Briggs, 2009; Hu & Trenkic, 2021; Powers & Rock, 1999). The effects reported were inconsistent. Briggs (2001) and Domingue and Briggs (2009) derived data from the National Educational Longitudinal Study in the United States to explore the relationship between students' use of different test preparation methods and the score gains on the SAT or ACT. The results showed that the use of a private tutor and enrollment in commercial coaching courses are the only test preparation methods that have small but positive effects on SAT-mathematics scores (about 13–15 points); while no forms of test preparation had statistically significant effects on SAT-verbal scores and ACT scores.

Studies conducted in different test contexts reported small or even negative effects of instructor-driven test preparation (i.e., Griffin et al., 2013; Trenkic & Hu, 2021; Xie, 2013). For example, Xie (2013) examined the test preparation effect on scores of College English Test, and the multiple regression and structural equation modelling results revealed that preparation for the test by drilling did have a significant effect on improving test scores (score gains .27 SDs), while other test preparation approaches such as developing general language skills did

not affect test scores. Griffin et al. (2013) investigated whether attending a commercial test preparation course affects performance in the Undergraduate Medical and Health Sciences Admission Test (UMAT). They found that enrollment in commercial coaching courses had inconsistent effects on students' performance in different sections of the test and for different levels of student ability. Their results showed that test preparation had no effect on student performance in the problem-solving and "understanding people" sections of the test, while it had positive effects for high ability students on their performance in nonverbal or abstract reasoning and negative effects for lower ability students.

Given the inconsistent findings of the aforementioned studies, it would be problematic to draw a conclusion about the test preparation effect. Besides, some studies using self-report surveys lacked control of confounding variables and ignored the heterogeneity of test preparation (Briggs, 2009). Thus, the findings from these studies may have limitations in accurately capturing the effect of test preparation. Besides, specific characteristics of test preparation activities, such as material use, duration, and students' strategy use, etc., were rarely collected and reported in most self-report surveys. As a result, empirical research has so far been unable to provide a comprehensive picture of test preparation activities and their effects; thus, a research synthesis is required.

Previous Reviews

Tracing back to the 1980s and 1990s, a considerable number of review articles synthesizing (quasi) experimental studies on test preparation effects were published primarily on coaching, but the effect reported in these articles varied. One of the earliest and influential meta-analytic reviews conducted by Bangert-Drowns, Kulik, and Kulik (1983) investigated the coaching effect on achievement tests using data from 30 controlled studies. This meta-analysis review reported a significant effect of coaching programs on ACT scores (average ES = .25), which was affected by the level of training intervention. Specifically, they found that test preparation programs designed to improve broad cognitive skills yielded the largest effects (ES = .66), and extensive programs, which included drill practice, had medium effects (ES = .43), while short test-oriented sessions had smaller effects (ES = .17). Bangert-Drowns et al. argued that their finding was in line with what exam boards had long claimed—that long-term preparation on broad knowledge and abilities could have a larger effect than simple drill or practice on items. However, there were some inherent limitations to this research due to the imbalance in the number of studies included, raising concerns about the generalizability of the findings. Most interventions (22 out of 30) included in this meta-analysis were short test-oriented programs, while only one included an intervention focused on broad skill training. The effect of coaching of broad skills was highly likely to be overestimated by pooling the effect size from only one study.

In the meta-analysis they published later on the effect of coaching on aptitude tests, Kuilk, Bangert-Drowns, and Kulik (1984) found that aptitude tests such as SAT and intelligence tests were less affected by coaching than other aptitude tests. Coaching on the SAT had a small effect (estimate ES = .15), whereas coaching on other aptitude and intelligence tests had a substantial effect (estimate ES = .43). This finding was echoed by Becker's (1990) comprehensive synthesis of the

effect of coaching programs on SAT scores, in which she concluded that although test preparation improved SAT scores in general, its effect was small and varied widely in terms of test domain. According to her study, test preparation interventions raised SAT-verbal scores by .09 SDs and SAT-math scores by .16 SDs, which implied that the math section in aptitude tests is more susceptible to test preparation than the verbal section.

Similar results were reported in a more recent systematic review of the effects of preparatory courses on SAT scores (Montgomery & Lilly, 2012). Ten experimental studies published between 1970 and 2000 related to the effect of SAT preparation courses yielded an average of 23.5 points of score gain on the verbal subtest (800 points in total) and 32.7 points gain on the math subtest (800 points in total) separately. Moreover, long coaching programs (more than 8 hours) posted a significantly larger effect on SAT-math scores than short programs (8 hours or less).

In addition to reviews on coaching effects, there were also some reviews focused on the effect of taking tests repeatedly on score increases, which is known as the practice test effect or the retest effect. The earliest meta-analysis of practice effect was conducted by Kulik, Kulik, and Bangert-Drowns (1984). Their study, which synthesized data from 40 studies related to the effect of taking practice forms of the test prior to the criterion aptitude or achievement test, revealed that taking practice tests that have identical form with the criterion test yielded a significantly larger effect on test scores (ES = .42), compared with practice tests that used alternative forms (ES = .23). Kulik, Kulik, and Bangert-Drowns (1984) finding is broadly in line with two later meta-analyses on the retest effect for cognitive ability tests, conducted by Hausknecht et al. (2007) and Scharfen et al. (2018). Hausknecht et al.'s (2007) study reported an adjusted overall effect size of .26, while Scharfen et al.'s (2018) study revealed that retaking cognitive ability tests multiple times can lead to score gains of .5 SDs, and they also found that the retest effect stagnates after taking the test more than three times. However, considering that it is a gradual process for students to learn domain-specific knowledge and develop relevant abilities, whether practice tests on their own can increase test scores on proficiency tests or knowledge tests is still questionable and needs further exploration.

To sum up, previous reviews identified a small coaching effect and practice effect on test performance. However, there were several limitations in these reviews, making it difficult to generalize the findings. To be specific, previous reviews on test preparation effects concentrated on the coaching effect for college admission tests in the US context or the practice effect on cognitive ability tests, whereas the effect of test preparation on other proficiency tests or knowledge tests of a specific content domain in a wide range of educational testing contexts was not addressed by existing reviews, thus impeding our understanding of the variance of test preparation effect.

Moreover, most reviews were published decades ago and featured test preparation effect from articles published before the year of 2000. Findings drawn from these articles may not be applicable in today's educational environment. For these reasons, this meta-analysis is conducted to provide more up-to-date conclusions about test preparation effects and expand results to multiple educational contexts.

Theoretical Underpinning of Test Preparation Effect

To explain the effect of test preparation, various interpretations have been made in previous empirical studies and reviews, such as the improvement of memory and knowledge retention (e.g., Hausknecht et al., 2007), the development of proficiency or abilities (e.g., Messick, 1982; Xie, 2013), the increase of test-wiseness (e.g., Madaus, 1988), the increase in motivation and confidence (e.g., Hong & Peng, 2008; Xie & Andrew, 2013), or the reduction of test anxiety (e.g., Messick & Jungeblut, 1981). Arendasy et al. (2016) summarized four possible models to explain the causes of test preparation effects from existing literature. The first postulates that score gains should be attributed to the familiarization of test-taking and test structure, which reduces construct-irrelevant variance. For test-takers, the first few test questions are easier because they do not need to familiarize themselves with the structure of the test. However, such gains caused by familiarization cannot reveal students' abilities because they are not relevant to the construct being assessed. The second model, according to Arendasy et al., states that engaging in test preparation improves test-wiseness and test-taking skills, such as guessing and eliminating wrong answers, which can improve the likelihood to correctly solve test items, without affecting the underlying construct. The third model ascribes the score gain to teaching construct-relevant materials, which further results in the development of test-specific cognitive abilities or an increase in domain-specific knowledge. The fourth model is based upon the premise that the increase in test scores is caused by the improvement of generalizable cognitive abilities.

These four models form a comprehensive theoretical framework of test preparation effect. There are numerous potential factors influencing the effect of test preparation that can be deduced from the four models, which are investigated in the current paper. Moreover, given that Arendasy et al.'s (2016) explanations were mainly drawn from retest effect literature, we will examine the applicability of this framework in the test preparation context and explore other possible underlying mechanisms of test preparation effect.

Potential Moderator Variables

In the aforementioned empirical studies and reviews, it has been shown that instructional characteristics in test preparation intervention (e.g., methods, the use of practice tests, instructional emphasis, duration, transferability) contribute to the effect on test score improvement. In addition, a body of literature has indicated that test-specific features (e.g., domain, type of tasks, cognitive ability) and individual differences (e.g., age, ability level, prior attainment, previous test experience) explain variations in test preparation effects (e.g., Arendasy et al., 2016; Briggs, 2009; Kulik et al., 1984; Powers, 1986; Powers & Swinton, 1984). In line with previous empirical evidence, we examined whether, and the extent to which, different implementations of test preparation in different testing contexts have varied effects on individuals' test performance. Moreover, the effectiveness of test preparation is likely to be contingent on study design (Becker, 1990; Briggs, 2009; Hausknecht et al., 2007); thus, design characteristics were also examined in our moderator analysis.

Design characteristics of test preparation intervention

The following moderator variables have been selected as test preparation design characteristics: type of intervention, level of intervention, instruction on strategy use, material use, and duration of the intervention.

Type of intervention refers to the specific test preparation methods, which were organized into three basic categories, including school courses, commercial courses, and self-study relevant materials. The effect variance related to test preparation intervention offered by different institutes has not been well explored in previous meta-analyses.

We also analyzed the *breadth of intervention*, which comprised three levels based on the instructional and practical emphasis: 1) broad on a general knowledge/skill area or test-associated curriculum, without a narrow focus on the test; 2) specific instruction on test-specific content; and 3) narrow (or drilling) instruction on exactly the same tested items. The level of instruction might explain the variance of the effects between the studies. Bangert-Drowns et al. (1983) found in their meta-analysis that coaching programs designed to improve broad cognitive abilities yielded the largest effect. In contrast, some empirical studies showed that specific test preparation focusing on test-related content or drilling on tested items has significant effects on test scores, although the predictive validity was not improved in this way (Anastasi, 1981; Allalouf & Ben-Shakhar, 1998; Arendasy et al., 2016; Hu & Trenkic, 2021; Xie, 2013).

Moreover, we examined whether the *instruction on strategy use* is a potential moderator variable. It has been found that the strategies students use during test preparation are linked to test performance (Purpura, 1999; Phakiti, 2003). Possible effective strategies include test preparation management, test-taking skills, memorization, socio-affective strategies, and broad knowledge/skills development strategies (Hsiao & Oxford, 2002; Oxford, 1991; Xie, 2013). Xie (2013) examined the effectiveness of test preparation strategies, finding that drilling and test-taking skills were the most effective ways of affecting test scores, although the effect was small.

Furthermore, we investigated if different *material use* in interventions could lead to the variance in test preparation effect. The effect of varied use of materials has rarely been examined by previous meta-analyses, and we expect to identify the kinds of materials that could improve test performance effectively. The use of practice tests, sample items, workbooks, and computer programs was investigated to see if the test preparation effect would be moderated by these materials. Finally, we included *test preparation time* of the intervention to see if the amount of contact hours and overall duration of the intervention (weeks) might explain the variance of the effects between studies. Bangert-Drowns et al. (1983) found that the duration of the coaching program positively correlates with the magnitude of the coaching effect.

Test-specific characteristics

The following two moderating variables associated with test-specific features were seen as relevant: academic domain and nature of the test.

We examined whether the *academic domain* assessed by the test moderated the test preparation effect. Research indicated that tests consisting of analytical or

quantitative tasks were more coachable and benefited more from repeated test-taking than did verbal tasks (Hausknecht et al., 2007; Kulik et al., 1984). We coded the subject domain to investigate whether the variance of test preparation effects was affected by domain type.

Moreover, we included *nature of the test* in the moderator analysis to see whether test preparation was as effective for high-stakes tests as it was for low-stakes tests, for curriculum-based tests as it was for non-curriculum-based tests, and for college admission tests and other types of tests. Only very limited numbers of studies examined test preparation effects on low-stakes tests (e.g., Brunner et al., 2007); thus, previous reviews mainly focused on the test preparation effect on high-stakes tests, especially college admission tests.

Student characteristics

There is plentiful evidence that individual differences are possible moderators of students' test preparation effect (Appelrouth et al., 2017; Xie & Andrew, 2013). For example, Kulik et al. (1984) found that the average test preparation effect is highest for high-ability students (ES = .82). Similarly, Griffin et al. (2013) found that coaching had a significant effect on the reasoning section of UMAT for high ability students, while for those of lower ability, the effect was negative.

Besides, it has been indicated that students' past test experience might moderate test preparation effect. Domingue and Briggs's (2009) study revealed that only students with previous test-taking experience can significantly benefit from private tutoring on the ACT.

In addition, test preparation processes could also operate differently for students from different education stages and age groups due to differences in self-regulated learning abilities, which tend to increase at higher ages. Kitsantas (2002) argued that students who use more self-regulatory skills (such as goal setting, keeping records and monitoring, and self-evaluation) during preparation have higher test performance.

We therefore investigated whether student characteristics such as *education level* and *previous experience of test-taking* influence the degree to which test preparation can be used to effectively improve test performance and long-term development.

Study design

It is also possible that studies with a more rigorous experimental study design yielded greater effects than studies with lower quality. Previous review articles indicated that randomized experiments reported significantly higher effects of test preparation on test success than quasi-experiments (Kulik et al., 1983, 1984); however, only a small proportion of studies included in previous reviews used a randomized design with a control group (Bangert-Drowns, Kulik, and Kulik, 1983; Briggs, 2009). Thus, in this study, we also examined whether *allocation type* and *control group type* are possible moderators that predicted the magnitude and direction of the test preparation effect. To note that control conditions in most studies are some forms of "business as usual" (e.g., Alderman & Powers, 1980; Bookman, 1981), while some studies compare two types of test preparation

practice—for example, intensive test preparation vs. instructions on general skills or broad knowledge (e.g., Nishitani, 2006; Robb& Ercanbrack, 1999).

Method

Literature Search

The literature search was initially carried out in December 2019 using five electronic bibliographic databases: *ERIC*, *ProQuest*, *PsycInfo*, *Web of Science*, and *GoogleScholar*. All the databases were searched simultaneously. In January 2024, the first author checked these databases again to locate newly published relevant studies. To access the initial body of relevant literature, the publication date was not specified. Key- and subject-word searches were conducted to capture all possible literature. The standard truncation (*) was also used for retrieving variations of the search terms. Searches utilized the following words and word combinations: all (test prep* OR test preparation OR shadow education* OR coaching OR private tutor* OR tutoring) AND all (test* OR test-tak*) AND all (test performance OR test score* or achievement* OR attainment*) OR all (effect OR washback* OR practice effect*). In addition, a secondary literature search was conducted by scanning reference lists from recent meta-analytic investigations of test preparation effects and practice effects on educational tests (i.e., Adesope et al., 2017; Hausknecht et al., 2007). The Social Science Citation Index database was also used to search for journal articles that referenced these meta-analyses. Grey literature that may be applicable to this study, such as dissertations, conference papers, and reports, was also included. After the literature screening, we conducted a backwards search to scan reference lists of all included studies, as well as a forward search to scan articles that cited the included studies. A first author search was also conducted to determine if authors of the included studies published other relevant articles. The previously mentioned electronic and manual literature covered all of the papers that were available through 2024.

Inclusion and Exclusion Criteria

The inclusion and exclusion criteria of this meta-analysis were:

1) The included articles examined the effect of test preparation activities as listed in the working definition. The treatment condition(s) had to include any form of test preparation intervention (i.e., test preparation courses or computer programs) or instructional intervention toward test-taking for one specific test. Articles focused on other educational activities, whose main purpose was not for test preparation, even though they used test scores to indicate students' attainment, were excluded (e.g., Hensley et al., 2021).
2) The included articles investigated test performance of large-scale tests designed for educational purposes and conducted in real-life educational contexts. Articles investigated the preparation effect on course tests or tests developed by researchers (e.g., Bol & Hacker, 2001), such as memorization tests, cognitive tests in operational contexts, and neuropsychological tests administered for clinical diagnoses, were excluded.

3) The included articles examined test preparation effect on a non-self-reported measure of test performance. Articles used self-reported test scores as dependent variables were excluded.
4) The included articles used either an experimental or quasi-experimental design with at least one treatment group and one control group, which allowed drawing causal inferences regarding the effect of test preparation on test performance. Initial equivalence between the treatment and the control groups must be assured. Included studies must provide pretest measures or covariate-adjusted measures from which we could compute pretest- and/or covariate-adjusted effect sizes (Ye et al., 2023). Studies without a control condition and/or studies without initial equivalence between the treatment and the control were excluded (e.g., Gaye, 2001; Lane, 2009). In addition, articles that reported effects that cannot be attributed to test preparation intervention were excluded. For example, some articles asked students to report whether they participated in test preparation or not and made comparisons among their scores (e.g., Li & Xiong, 2018; Xie, 2013).
5) Sufficient data (i.e., means, standard deviations, sample sizes) should be reported in each included study to enable the computation of effect sizes.
6) Sufficient information about the test preparation programs should be reported, such as the materials utilized, the teaching methods employed, the learning activities students undertook, the duration of the program, etc. Articles without a detailed description of the test preparation procedures were excluded.
7) Included studies could be conducted at any education level and were not limited to any specific study site, but only studies published in English were included to ensure the reliability and coherence of coding the key elements of test preparation intervention.
8) Only SAT-related studies conducted after 1978 and ACT-related studies conducted after 1985 were included, as example items of the two tests were not provided by the testing bodies to the public prior to the two years (A [Mostly] Brief History Of The SAT And ACT Tests, n.d.), so teachers and students had very limited access to test-related materials. Any SAT-related studies conducted before 1978 and any ACT-related studies conducted before 1985 were excluded.

Literature Filtering

Of the 5629 retrieved records in total, 2228 duplicates were removed. We then screened titles, abstracts, keywords, and subheadings of each article for possible inclusion. The first author independently screened all articles, and a trained research assistant screened 50% of the articles. For these studies in which the abstract did not provide sufficient information, we read sections related to the research design to gather more information. Studies not examining issues related to test preparation effect or not using quantitative methods were excluded. The interrater reliability for the title-abstract screening was .92, reflecting a good level of agreement. The disagreements were solved by discussion. After the screening

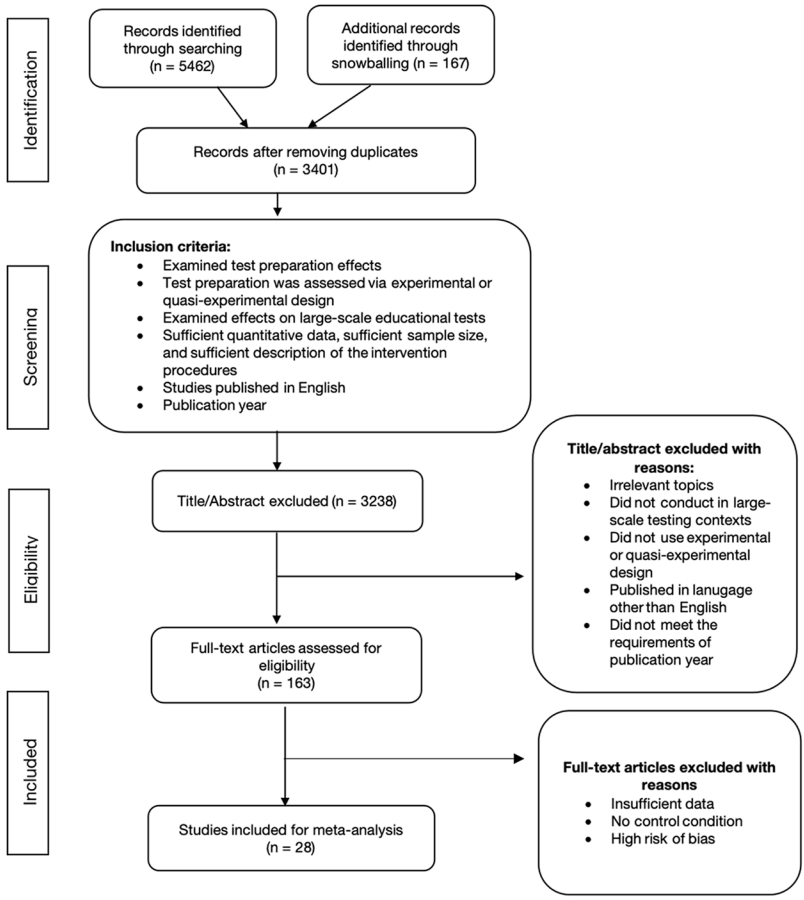

FIGURE 1. *Flow diagram of literature search and processing of records.*

stage, 163 studies were retained. The first author and the trained research assistant engaged in full-text screening independently, reading the full texts of all 163 studies to determine whether they met all inclusion criteria. The interrater reliability for full-text screening was .87. At this stage, disagreements primarily centered around inclusion criteria 4, 5, and 6. For example, some studies reported themselves as experimental studies but lacked a complete control condition, or their effect could not be attributed to the test preparation intervention because covariates were not controlled during the data analysis. These issues required careful reading and discernment by the raters, which posed challenges in terms of judgment. The two raters discussed their disagreements until a consensus was reached. As a result, a total of 28 studies met all the inclusion criteria and were included in the subsequent analysis. The literature searching and processing stages are shown in Figure 1.

Data Extraction

The included studies were systematically analyzed by making use of two coding schemes: one on the study level and another on the effect size level. In the first scheme, five groups of variables were coded for each study, including general study descriptors (e.g., study type and publication year), test preparation characteristics (e.g., duration of intervention, type of intervention, level of instruction), student characteristics (e.g., gender, ability level, education level), test-related variables (e.g., subject domain, nature of test), and study characteristics (e.g., type of comparison group, allocation). In the second coding scheme, raw data were collected to calculate the particular effect size, and all information about the dependent variable (the test scores, or assessment on long-term retention) was gathered.

We took an iterative approach to extract data whereby three coders refined the classification of each variable as they progressed through the included studies during coding trials to ensure that the classifications best characterized the literature. The utilized classification of each coding variable is listed here.

Nature of publication.

Publications were classified as either journal articles, dissertations, or reports (conference reports, industrial reports, or exam reports). The year of each publication was specified.

Country

The country in which the study was conducted was specified.

Nature of test

The nature of the test was classified as either curriculum-based or non-curriculum-based. Taking the two widely used college admission tests in the United States as examples, ACT is a curriculum-based test while SAT is a non-curriculum-based test.

Stakes of the test

The stakes of the test that students prepared for were classified as either high stakes or low stakes. High-stakes tests refer to tests with important consequences for students (Madaus, 1988), while low-stakes tests do not have important consequences associated with test performance. It is important to note that whether the test is high-stakes or low-stakes is relative to test-takers. For example, in one of the included studies, Hardison and Sackett (2008) examined the effect of short-term, rule-based coaching on standardized writing tests using writing prompts from the College Level Examination Program (CLEP) to examine students' writing performance. The CLEP can be high stakes for students who want to use it to receive college credit. However, in Hardison and Sackett's (2008) study, students were not attending the real CLEP but used CLEP prompts to assess their writing improvement after the intervention; thus, the test in this study was coded as low stakes.

Test type

Based on the evaluation of all included articles, we classify the large-scale educational tests in the meta-analysis as either college admission tests (such as SAT, ACT) or others.

Subject domain of the test

The subject domain focused by the target test in each individual study was classified as either verbal, quantitative/mathematics, science, foreign language, or mixed. If outcomes from multiple domains were reported in the article, then the outcome on each domain were recorded and separate effect sizes were reported. If the study reported an overall combined score of multiple domains, then the subject was coded as "mixed."

Format of the test

The inclusion of multiple-choice questions in the test was classified as yes or no, as was the inclusion of short-response questions. In some included articles, authors did not mention information about the test format; thus, the coders checked the official website of the test to gather the information.

Test administration

The administration of the test was classified as either mock, indicating that mock tests were used for the posttest, or real, indicating that participants in the study sat for the real test after test preparation intervention.

Type of test preparation intervention

The test preparation intervention was classified as commercial courses, in-school courses, after-school courses, or self-study. Note that except for specialized courses provided by commercial test preparation centers, we regarded paid courses offered in public or private schools as commercial courses as well, because those paid courses were selective and some were taught by teachers from testing centers or test preparation institutes (e.g., Filizola, 2008). Courses provided by schools or universities were coded as either in-school courses or after-school courses (i.e., extracurricular activities).

Breadth of test preparation intervention

Based on previous studies (Bangert-Drowns et al., 1983; Kulik et al., 1984; Hill et al., 2008), we classified the breadth of test preparation intervention as broad, specific, or narrow. When the intervention focused on a general knowledge/skill area (but may have some overlap subjects with the tested content), it was coded as broad. When the intervention targeted the same subject or area as the test, but did not drill on the test, it was coded as specific. When the intervention only focused on the test-specific content and format or drilled on the exact same tested items, it was coded as narrow.

Strategy use

The provision of inclusion in each of the five test preparation strategies was classified as yes or no. The five strategies investigated by the current meta-analysis were adapted from Xie and Andrew's (2013) study:

- Test preparation management (TPM) strategy, which resembles metacognitive strategies and refers to test preparation practices through analyzing test papers to identify frequently assessed areas, learning marking rubrics and sample responses to evaluate one's own answers, conducting self-assessment of personal strengths and weaknesses, and effectively managing time for one's own test preparation process;
- Test-taking strategy (TTS), which refers to the practices that test takers explore and practice test-taking skills for different sections of the test, including guessing, context clues, eliminating incorrect answers, taking notes in the margin, and time management for test-taking, etc.;
- Drill, which refers to the intensive and repetitive practice of a narrow range of skills and knowledge tested by the test, such as memorizing frequently tested knowledge or facts and repeatedly practicing past test questions and remembering exemplar responses;
- Socio-affective (SOAF) strategy, which refers to test-takers' use of social strategies to seek support from teachers and peers, and their use of affective strategies to motivate themselves and to reduce test anxiety;
- Broad knowledge skills development (BKSD) strategy, which refers to the learning strategies that test takers use to develop broad knowledge or skills via extensive exposure to and functional uses of knowledge or skills in authentic contexts.

If relevant features of any of the previous strategies were described in a study, then the study was coded as "yes" for the use of the particular strategy. However, if no relevant feature was mentioned in any of the previous strategies, then the strategy was deemed not to have been used in the study and was coded as "no."

Material use

The use of each of the three types of materials was classified as yes or no. Materials included a commercial test preparation workbook, a timed practice test, sample items, and computer-based programs. If an included study did not mention the use of a specific material, we assumed it did not use the material and coded it as "no."

Time and duration of the intervention

We recorded the time (indicated by the number of contact hours) and duration (indicated by weeks) of each intervention.

Education level

Participants' education level was classified as either tertiary, secondary, primary, or mixed. If the study invited participants from a wide population and mixed with their educational levels, it should be coded as mixed (see Farnsworth, 2013, as an example).

School type

For included studies conducted in primary and secondary settings, we further classified two types of schools that provided test preparation courses: public school or private school. Some experiments did not provide enough description for the school type, thus they were coded as "unknown."

Previous test experience

We recorded participants' previous test experience. Four categories of participants' experience were found in the included studies: yes, no, parts of students have relevant experience, or not mention.

Allocation

Participant allocation to condition was classified as random allocation or nonrandom.

Sample size

Studies with small sample sizes might report larger, positive effect sizes than studies with larger sample sizes (Cheung & Slavin, 2016; Slavin & Smith, 2009). Thus, we recorded the sample size of each study and evaluated whether it served as a significant moderator. Additionally, the sample size was controlled as a covariate in the analysis of other moderators.

Comparison group

We identified two types of comparison groups in the included studies: no-test-preparation condition and different test preparation conditions. Participants in a no-test-preparation condition usually did not receive any instruction or received instruction after the intervention (e.g., Moss et al., 2012). Some included studies compared the effects of different types of test preparation, which could be characterized as a business-as-usual condition in many instances; that is, two versions of a course were run—one provided specific test preparation tasks or instructions, and one provided the normal course as usual. For example, Winke and Lim (2017) compared the effects of three different interventions—explicit instruction, implicit instruction, and control—on second language listening test performance. We therefore extracted two comparisons from this study: explicit instruction vs. control and implicit instruction vs. control. The control condition provided students with instructions on American culture, which was regarded as broad instruction on second language learning, thereby coded as a different test preparation condition.

Coding and Interrater Reliability

Three rounds of coding training were performed to allow coders to discuss and refine the classification to characterize each individual article. Within the coder training, 20% of included studies ($N = 6$) were coded by all three coders (with an interrater agreement above .80 for each article). Disagreements were discussed and resolved, while the coding scheme was modified based on the discussion. After coding trials, the remaining articles were randomly grouped into two sets. The first author coded all included articles independently, and the other two coders coded 50% each, with random allocation. The initial interrater reliability was .89. The discrepancies were resolved through discussion.

Risk of Bias (RoB) Assessment

To ensure that the accuracy of the meta-analysis was not compromised by including study designs of varying quality, we conducted a comprehensive RoB assessment for all effect sizes individually. Since we amalgamated results from

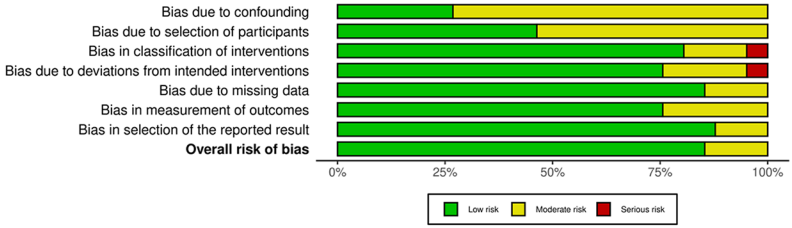

FIGURE 2. *ROBINS-I weighted summary plot.*

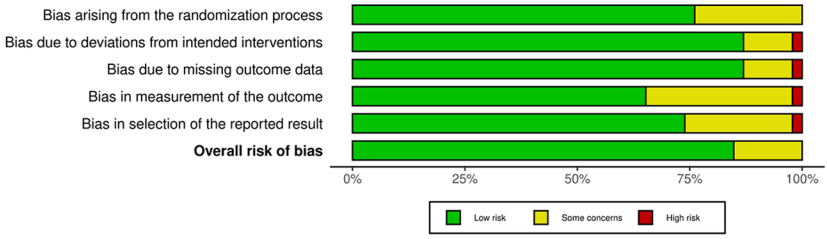

FIGURE 3. *RoB 2 weighted summary plot.*

both randomized and nonrandomized studies, we applied the RoB 2 tool for RCTs (Sterne et al., 2019) and the ROBINS-I tool for nonrandomized studies (Sterne et al., 2016). The RoB assessment for all included studies was conducted by the first author and the trained research assistant. The interrater reliability between the two raters was .94. Disagreements were solved by discussion.

Figures 2 and 3 depict summary plot results of the RoB assessment for nonrandomized cases and randomized cases, respectively. There were 14 included studies that reported 41 cases that were nonrandomized, which were assessed via the ROBINS-I tool; while 15 included studies reported 46 randomized cases that were assessed via the RoB 2 tool. As shown in the two figures, the majority of the included cases across all research designs were assessed as having a low risk. The characteristics of the research design were further analyzed in the moderator analysis to illustrate whether varying features of the design influence the magnitude of the test preparation effect. In addition, a sensitivity analysis was performed by conducting the meta-regression models, retaining studies that were assessed as low risk of bias in all domains. The restricted model results were then compared with the full model.

Statistical Analytical Strategy

A correlated and hierarchical random-effects, meta-analysis model was conducted using R version 3.6.3 (R Core Team, 2019). For the analyses, we followed the procedure described by Borenstein et al. (2009) for effect size calculation,

integration, and meta-regression for moderator analysis. We used the standardized mean difference effect size to measure the effects of test preparation interventions between treatment and control groups on student test performance. The data on the effects of the included studies were gathered in an Excel sheet. Cohen's *d*, variance, and *SE* of Cohen's *d* were computed in Excel. Then, Hedges' *g* was calculated and effect aggregation was performed in R Studio using the "metafor" and "robumeta" packages. When both pretest and posttest data were available, pretest data were used to adjust the effect sizes by subtracting the pretest effect sizes from the posttest effect sizes and adding up the variance of both effects.

Many studies included in our meta-analysis either reported multiple outcomes using the same participant groups or nested several treatments and control conditions in one study. Including those dependent effect sizes can bias the results of the analysis, as studies reported more effect sizes would take a larger account. Therefore, we employed robust variance estimation (RVE) to handle dependent effect sizes (Tanner-Smith et al., 2016), which accounts for both hierarchical effect cases (i.e., multiple experiments are nested) and correlated effect cases (i.e., different measures with the same participants) with small-sample corrections.

Heterogeneity in the effect sizes was estimated using the Cochran Q-test. The I^2 statistic indicates that the proportion of the variation in effect sizes is due to true between-study heterogeneity rather than sampling error, and $I^2 > 75\%$ represents considerable heterogeneity (Higgins & Thompson, 2002). Meta-regressions and subgroup analysis were performed to examine the moderating effects of various factors that differed across studies.

To investigate the moderating effects of possible factors that varied across studies, we conducted both subgroup analyses and meta-regressions. Although subgroup analysis could recognize whether there were differences between subgroups by determining whether the confidence intervals around their effect sizes overlap, heteroscedasticity or multicollinearity may lead to biased estimates (Steel & Kammeyer-Mueller, 2002). To examine the overall impact of a moderator, we performed separate meta-regressions for each predictor, controlling the sample size as the covariate (Cheung & Slavin, 2016; Slavin & Smith, 2009).

Assessment of Publication Bias

Publication bias is one of the threats to the validity of meta-analyses (Van Aert et al., 2019). It occurs as a result of some studies not being published due to the lack of statistical significance in their results (Banks et al., 2012). Carter et al. (2019) suggested that if the probability of publication bias is high, the analysis should not rely only on the random effect model. For assessing and reporting publication bias, we conducted a moderator test between published and unpublished studies to explore the possibility that published studies had significantly higher effect sizes than unpublished studies. Funnel plots were also produced in R Studio to inspect asymmetry and heterogeneity. Considering that we have a relatively small sample of empirical studies, the funnel plots and Egger's regression test were used to evaluate whether small-study effects existed.

Results

Descriptive Statistics

Overall, findings from 28 primary studies reporting 92 effect sizes were extracted for the meta-analysis. Descriptive information of all included studies is reported in Table 1. The majority of studies investigated test preparation effects on college admission tests such as ACT and SAT ($n = 12$) and English proficiency tests such as IELTS ($n = 11$), while a small number of studies ($n = 5$) were conducted in other testing contexts. Most studies were conducted in the United States ($n = 15$), while seven studies were conducted in Asian countries and regions. The majority of interventions reported by the studies included ($n = 27$) were led by instructors, while only one study investigated students' self-preparation using computer programs and handouts prepared by teachers. Most of the effect sizes ($n = 87$) reported interventions that prepared students for one specific test and used the criterion test to measure students' performance. However, four studies utilized different tests to examine the transferability of test preparation effects and reported five relative effect sizes.

Overall Meta-analysis of the Effect of Test Preparation

In the overall model, a significant positive effect of test preparation on large-scale criterion test performance was found: $g = .26$, $se = .08$, 95% CI = .10–.42, $p < .001$. A forest plot of all recorded effect sizes is provided in Figure 4. According to Hattie (2008) and Kraft (2020), this effect can be considered large (greater than .20). When interpreting this effect size in practical terms, an effect size of .26 indicates that, if the data follow a normal distribution, a randomly selected individual from the experiment group has approximately a 60% probability of scoring higher than a randomly selected individual from the control group. This represents a meaningful improvement in the context of educational testing.

The heterogeneity between the studies' effect sizes was significant with $I^2 = 82.44\%$, $Q(86) = 331.79$, and $p < .001$, which shows that the test preparation interventions do not share the same true effect size and the heterogeneity is high (Higgins & Thompson, 2002). The SD of the true effect size is estimated at $\tau = .37$, and the variance of the true effect sizes is estimated at $\tau^2 = .13$. The prediction interval is .62–2.69. This supports the use of a meta-regression and subgroup analysis in order to explain the observed heterogeneity in effect sizes.

This meta-analysis also investigated the transferability of the test preparation effect. A small number of included studies ($n = 4$, $k = 5$) explored the transferability of preparing for a specific large-scale test to other tests within a similar domain, such as whether the IELTS preparation course could improve the performance of the Online Oxford Placement Test (OOPT) (Trenkic & Hu, 2021), or if the training on SAT writing skills benefited other writing tasks such as writing personal statements or complaint letters (Hardison & Sackett, 2008). The results showed that the effect of test preparation seems to be unidirectional to a specific test and cannot be transferred, as the effect size was small and not significant: $g = .06$, $s = .08$, $p = .502$. However, we should be very careful with the interpretation of this result, as our data were drawn from a limited number

TABLE 1

Descriptive characteristics of the included studies

Authors	Year	Type	Region	Test	Domain	Education Level	Number of Participants
Alderman & Powers	1980	Journal article	US	SAT	Verbal	Secondary	559
Allalouf & Shakhar	1988	Journal article	Israel	PET	Mixed	Secondary	274
Bailey & Judd	2018	Journal article	South Korea	TOEIC	English	Tertiary	65
Bookman	1981	Dissertation	US	Comprehensive Test of Basic Skills	Mathematics	Secondary	463
Brunner et al.	2007	Journal article	German	PISA	Reading & mathematics	Secondary	1323
Bunting & Mooney	2001	Journal article	UK	11-Plus	Mixed	Primary	552
Ercanbrack	1999	Journal article	Japan	TOEIC	English	Tertiary	488
Farnsworth	2013	Journal article	US	Oral English Tests	English	Mixed	34
Filizola	2008	Dissertation	US	SAT	Mixed	Secondary	24
Green	2007	Journal article	UK	IELTS	English	Tertiary	476
Hardison & Sackett	2008	Journal article	US	Writing	Verbal	Tertiary	92
Holmes & Keffer	1995	Journal article	US	SAT	Verbal	Secondary	70
Holthaus	2008	Dissertation	US	SAT	Mathematics	Secondary	33
Justus	2010	Dissertation	US	ACT	Mixed	Secondary	36
Lee	2019	Journal article	Taiwan	TOEIC	English	Tertiary	135
McMann	1994	Dissertation	US	ACT	Mathematics	Secondary	196
Mitchell et al.	2016	Journal article	US	DEA	Verbal	Primary	679
Moss et al.	2012	Exam board report	US	ACT	Mixed	Secondary	107
Nishitani	2006	Book chapter	Japan	TOEIC	English	Tertiary	102
Pan	2016	Journal article	Taiwan	GEPT	English	Tertiary	72
Parrott	2012	Dissertation	US	ACT	Mixed	Secondary	333
Rao	2003	Exam board report	Fiji	IELTS	English	Tertiary	71
Reynolds	1988	Journal article	US	PSAT	Mixed	Secondary	140
Schueler	2020	Report	Kenya	KCPE	Mixed	Primary	957
Takallou	2015	Journal	Iran	Konkoor	English	Secondary	260
Trenkic & Hu	2021	Journal article	Mainland China	IELTS	English	Tertiary	89
Winke & Lim	2017	Journal article	US	IELTS	English	Tertiary	63
Zuman	1987	Journal article	US	SAT	Mixed	Secondary	88

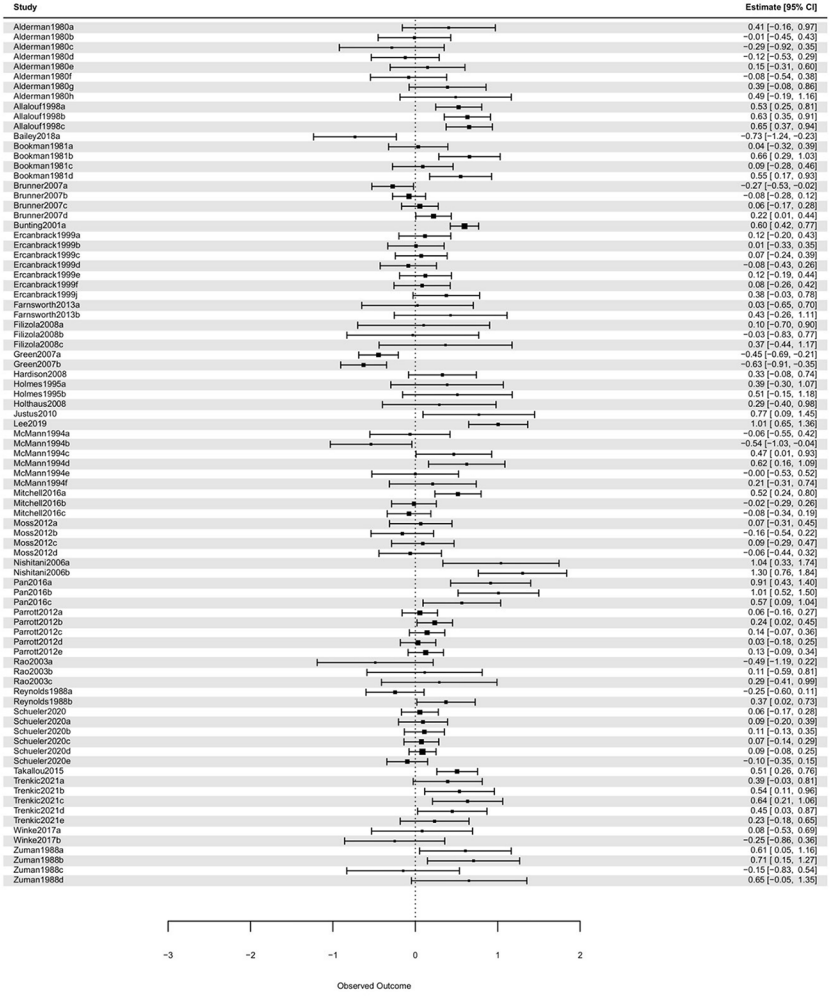

FIGURE 4. *Forest plot of all recorded effect sizes.*

of studies. Further research is needed to justify the transferability of the test preparation effect.

Moderator Analysis

We examined whether test-specific characteristics, test preparation–related factors, student-related factors, and study design characteristics explained the variance in effect sizes. Effect sizes for subgroups are presented in Table 2. Results of the meta-regressions are presented in Table 3.

TABLE 2

Results of the subgroup analysis

	N	*k*	Sample size	*g*	*SE*	I^2	*p*
Nature of the Test							
Test Stakes							
High-stakes tests	20	67	13243	.30	.08	81.81%	<.001
Low-stakes tests	9	20	3498	.16	.12	83.87%	.186
Relationship With Curriculum							
Curriculum-based tests	10	32	10282	.29	.12	84.76%	.012
Non-curriculum-based tests	18	55	6459	.30	.09	81.92%	.002
Type of Tests							
College admission tests	16	56	11805	.14	.09	77.40%	.156
Other tests	12	31	4936	.41	.11	81.17%	<.001
Subject Domain							
Verbal	12	27	4443	.22	.09	72.85%	.020
Mathematics	12	22	3514	.10	.10	68.51%	.006
EFL	12	30	4314	.28	.10	88.17%	.006
Others	3	4	2354	.15	.11	8.65%	.179
Mixed	4	4	2116	.29	.12	9.02%	.017
Test Format							
MCQ Questions Included							
Yes	23	73	14926	.30	.07	79.01%	<.001
No	7	14	1815	.08	.12	82.85%	.505
SR Questions Included							
Yes	4	15	6555	.29	.15	68.28%	.057
No	25	72	10186	.25	.08	82.86%	.002
Test Administration							
Real test conditions	9	27	9247	.21	.08	64.51%	.031
Mocks	21	66	7494	.28	.11	85.18%	.002
Type of Test Preparation Intervention							
After-school courses	7	25	8542	.27	.09	68.66%	.055
In-school courses	15	40	6012	.22	.13	88.26%	.034
Commercial courses	5	19	2084	.31	.14	62.23%	.040
Self-study	2	3	103	.38	.08	0%	.263
Breadth of Test Preparation Intervention							
Broad	7	14	2448	−.04	.13	73.31%	.761
Specific	10	26	8730	.33	.10	9.59%	.002
Narrow	17	47	5563	.32	.09	78.01%	<.001
Strategy Use for Test Preparation							
Test Preparation Management							
Yes	17	44	5759	.30	.10	83.28%	.003
No	12	43	10982	.21	.11	77.93%	.063

(continued)

TABLE 2 (CONTINUED)

	N	*k*	Sample size	*g*	*SE*	I^2	*p*
Test-Taking Skills							
Yes	21	62	7438	.35	.09	8.28%	<.001
No	10	25	9303	.05	.12	84.11%	.662
Memorization							
Yes	21	59	11007	.28	.08	81.66%	<.001
No	10	25	5734	.21	.09	8.11%	.019
Socio-affective Strategies							
Yes	10	24	6999	.41	.11	79.05%	<.001
No	22	63	9742	.20	.08	82.01%	.012
Broad Knowledge/Skill Development							
Yes	14	34	8435	.17	.09	81.22%	.069
No	20	53	8306	.31	.08	81.66%	<.001
Material Use for Test Preparation							
Practice Tests							
Yes	9	27	2704	.19	.13	78.94%	.233
No	20	60	14037	.29	.09	81.97%	.002
Sample Items							
Yes	25	71	15150	.23	.08	83.80%	<.001
No	5	16	1591	.10	.13	47.95%	.794
Commercial Workbook							
Yes	12	43	5788	.40	.10	75.65%	<.001
No	18	44	10953	.15	.08	83.49%	.077
Computer Program							
Yes	1	2	94	.47	.32	0%	.149
No	27	84	16647	.25	.08	83.04%	.001
Students' Education Level							
Primary	4	15	8080	.23	.19	91.22%	.220
Secondary	13	43	5472	.27	.11	6.89%	.019
Tertiary	11	29	3189	.25	.13	88.23%	.046
School Type							
Private school	3	8	383	.43	.18	12.93%	.021
Public school	9	31	5534	.22	.13	79.58%	.104
Commercial institute	4	15	1908	.27	.20	67.64%	.190
University	9	21	2531	.26	.14	9.17%	.075
Others	5	12	6385	.26	.19	73.54%	.185
Students' Previous Test or Test Preparation Experience							
Yes	4	6	995	.69	.19	6.55%	<.001
No	18	58	7335	.06	.08	79.35%	.477
Unknown	8	23	8411	.50	.11	75.03%	<.001

(continued)

TABLE 2 (CONTINUED)

	N	*k*	Sample size	*g*	*SE*	I^2	*p*
Random Allocation							
Yes	15	46	6017	.36	.09	73.49%	<.001
No	14	41	10724	.15	.10	78.71%	.128
Comparison							
No test preparation condition	20	65	14007	.30	.08	77.00%	<.001
Different test preparation conditions	9	22	2734	.14	.12	88.04%	.224
Publication Type							
Journal	18	52	7847	.23	.09	86.26%	.018
Dissertation	6	20	2661	.26	.17	42.52%	.131
Others	4	15	6233	.39	.20	8.78%	.050

TABLE 3

Results of meta-regressions

	b	*SE*	95% CI Low	95% CI Upper	*p*
Nature of the Test					
Test Stakes F (df1 = 2, df2 = 84) = .71, *p* = .497					
Intercept	.34	.11	.13	.55	.002
Low-stakes tests	−.14	.14	−.42	.13	.305
Sample size	−.00	.00	−.00	.00	.563
Curriculum F (df1 = 2, df2 = 84) = .31, *p* = .732					
Intercept	.31	.11	.10	.53	.004
Curriculum-based tests	−.09	.17	−.43	.24	.578
Sample size	−.00	.00	−.00	.00	.756
Type of Tests F (df1 = 2, df2 = 84) = 2.03, *p* = .138					
Intercept	.45	.12	.20	.69	<.001
College admission tests	−.27	.14	−.55	.01	.059
Sample size	−.00	.00	−.00	.00	.570
Subject Domain F (df1 = 5, df2 = 81) = .40, *p* = .851					
Intercept	.31	.12	.07	.54	.011
Verbal	−.05	.13	−.30	.21	.709
Mathematics	.01	.13	−.25	.27	.940
Others	−.10	.15	−.39	.19	.510
Mixed	.04	.16	−.27	.35	.781
Sample size	−.00	.00	−.00	.00	.538

(continued)

TABLE 3 (CONTINUED)

	b	*SE*	95% CI Low	95% CI Upper	*p*
Test format					
Included MCQ Questions F (df1 = 2, df2 = 84) = 1.65, *p* = .199					
Intercept	.13	.14	−.15	.40	.364
Yes	.23	.13	−.04	.49	.094
Sample size	−.00	.00	−.00	.00	.426
Included SR Questions F (df1 = 2, df2 = 84) = .25, *p* = .782					
Intercept	.29	.10	.09	.49	.004
Yes	.07	.16	−.25	.385	.683
Sample size	−.00	.00	−.00	.00	.519
Test Administration F (df1 = 2, df2 = 84) = .23, *p* = .797					
Intercept	.30	.10	.10	.51	.004
Real conditions	−.05	.13	−.32	.22	.722
Sample size	−.00	.00	−.00	.00	.625
Type of Test Preparation Intervention F (df1 = 4, df2 = 82) = .15, *p* = .963					
Intercept	.32	.17	−.03	.66	.070
In-school courses	−.06	.16	−.39	.26	.696
Commercial	.01	.23	−.45	.48	.954
Self-study	.07	.38	−.69	.83	.855
Sample size	−.00	.00	−.00	.00	.633
Level of Test Preparation Intervention F (df1 = 3, df2 = 83) = 2.69, *p* = .052					
Intercept	.00	.14	−.28	.29	.985
Specific	.39	.16	.07	.71	.016
Narrow	.37	.14	.10	.64	.008
Sample size	−.00	.00	−.00	.00	.457
Strategy Use for Test Preparation					
Test preparation Management F (df1 = 2, df2 = 84) = .28, *p* = .756					
Intercept	.25	.14	−.02	.52	.073
Yes	.07	.15	−.22	.37	.625
Sample size	−.00	.00	−.00	.00	.624
Test-Taking Skills F (df1 = 2, df2 = 84) = 3.08, *p* = .051					
Intercept	.03	.15	−.27	.33	.830
Yes	.31	.13	.06	.56	.017
Sample size	−.00	.00	−.00	.00	.871
Memorization F (df1 = 2, df2 = 84) = .31, *p* = .733					
Intercept	.26	.12	.03	.49	.027
Yes	.05	.08	−.12	.21	.579
Sample size	−.00	.00	−.00	.00	.599
Socio-affective Strategies F (df1 = 2, df2 = 84) = 2.10, *p* = .109					
Intercept	.24	.10	.04	.44	.017
Yes	.21	.11	−.00	.42	.047

(continued)

TABLE 3 (CONTINUED)

	b	*SE*	95% CI Low	95% CI Upper	*p*
Sample size	−.00	.00	−.00	.00	.502
Broad Knowledge/Skill Development F (df1 = 2, df2 = 84) = 1.27, *p* = .286					
Intercept	.34	.11	.13	.55	.002
Yes	−.13	.09	−.31	.04	.138
Sample size	−.00	.00	−.00	.00	.580
Material Use for Test Preparation					
Workbook F (df1 = 2, df2 = 84) = 2.25, *p* = .112					
Intercept	.17	.11	−.05	.40	.129
Yes	.24	.12	.01	.47	.044
Sample size	−.00	.00	−.00	.00	.797
Practice Tests F (df1 = 2, df2 = 84) = .47, *p* = .625					
Intercept	.34	.12	.11	.57	.004
Yes	−.12	.16	−.43	.19	.444
Sample size	−.00	.00	−.00	.00	.463
Sample items F (df1 = 2, df2 = 84) = 1.61, *p* = .107					
Intercept	.07	.14	−.20	.35	.593
Yes	.16	.12	.02	.41	.069
Sample size	−.00	.00	−.00	.00	.481
Computer Programs F (df1 = 2, df2 = 84) = .35, *p* = .707					
Intercept	.28	.10	. 08	.48	.007
Yes	.20	.33	−.46	.86	.546
Sample size	−.00	.00	−.00	.00	.624
Time Use for Test Preparation					
Contact Hours F (df1 = 2, df2 = 84) = .43, *p* = .652					
Intercept	.31	.10	.11	.50	.002
Contact hours	−.00	.00	−.00	.00	.489
Sample size	−.00	.00	−.00	.00	.618
Duration F (df1 = 2, df2 = 84) = .67, *p* = .517					
Intercept	.21	.13	−.05	.47	.111
Duration	.01	.01	−.01	.03	.314
Sample size	−.00	.00	−.00	.00	.625
Students' Education Level F (df1 = 3, df2 = 83) = .12, *p* = .946					
Intercept	.37	.29	−.22	.95	.216
Secondary	−.06	.27	−.60	.48	.829
Tertiary	−.08	.29	−.65	.49	.777
Sample size	−.00	.00	−.00	.00	.559
School Type F (df1 = 5, df2 = 81) = .40, *p* = .851					
Intercept	.46	.20	.07	.85	.022
Public school	−.21	.16	−.53	.11	.195
Commercial institute	−.17	.28	−.72	.38	.542

(continued)

TABLE 3 (CONTINUED)

	b	*SE*	95% CI Low	95% CI Upper	*p*
University	−.18	.24	−.65	.29	.448
Others	−.15	.28	−.70	.40	.595
Sample size	−.00	.00	−.00	.00	.694
Students' Previous Test or Test Preparation Experience F (df1 = 3, df2 = 83) = 8.19, $p < .001$					
Intercept	.14	.09	−.05	.33	.144
Yes	.66	.20	.27	1.05	.001
Not mention	.47	.11	.24	.69	<.001
Sample size	−.00	.00	−.00	.00	.134
Random Allocation F (df1 = 2, df2 = 84) = 1.36, $p = .263$					
Intercept	.18	.12	−.06	.42	.133
Yes	.20	.13	−.06	.46	.135
Sample size	−.00	.00	−.00	.00	.639
Comparison F (df1 = 2, df2 = 84) = .98, $p = .380$					
Intercept	.18	.13	−.08	.44	.165
No test preparation	.17	.13	−.09	.43	.205
Sample size	−.00	.00	−.00	.00	.492
Publication Type F (df1 = 3, df2 = 83) = .40, $p = .753$					
Intercept	.29	.18	−.06	.64	.098
Journal	−.02	.20	−.41	.37	.926
Others	.20	.27	−.35	.74	.477
Sample size	−.00	.00	−.00	.00	.447
Publication Year F (df1 = 2, df2 = 84) = .16, $p = .850$					
Intercept	−1.26	13.20	−27.51	24.99	.924
Year	.00	.01	−.01	.01	.907
Sample size	−.00	.00	−.00	.00	.569
Sample Size F (df1 = 1, df2 = 85) = .34, p = .560					
Intercept	.29	.10	.10	.49	.004
Year	−.00	.00	−.00	.00	.560

Test-specific characteristics

Nature and type of the test. According to the subgroup analysis, test preparation had a significant effect on high-stakes tests ($g = .30$, $se = .08$, $k = 67$, $p < .001$). The test preparation effect for low-stakes tests was not significant ($g = .16$, $se = .12$, $k = 20$, $p = .186$). According to the meta-regression, there was no significant difference between the test preparation effect for high-stakes tests and low-stakes tests, $b = -.14$, 95% *CI*: −.42 to .13, $p = .305$. Similarly, no statistically significant difference was identified between test preparation effect for non-curriculum-based tests (such as SAT and IELTS), $g = .30$, $se = .09$, $k = 55$, $p = .002$, and

for curriculum-based achievement tests (such as ACT), $g = .29$, $se = .12$, $k = 32$, $p = .012$; $b = -.09$, $se = .17$, 95% CI: −.43 to .24, $p = .578$. Regarding the type of test, the meta-regression showed that test preparation interventions on admission tests ($g = .14$, $se = .09$, $k = 56$, $p = .156$) and on other tests ($g = .41$, $se = .11$, $k = 31$, $p < .001$) did not yield significantly different effects, $b = -.27$, $se = .14$, 95% CI: −.55 to .01, $p = .059$.

Subject domain. The average test preparation effects per subject domain have been listed in Table 2. The results revealed small differences between subject domains. The meta-regression results showed that compared with EFL tests ($g = .28$, $se = .10$, $k = 30$, $p = .006$), the preparation for mathematics and verbal tests had smaller effects ($g = .10$, $se = .10$, $k = 22$, $p = .006$; and $g = .22$, $se = .09$, $k = 27$, $p = .020$), but the differences were not significant: $b = .01$, $se = .13$, 95% CI: −.25 to .27, $p = .940$; and $b = -.05$, $se = .13$, 95% CI: −.30 to .21, $p = .709$. There was no significant difference between other domains and EFL (see Table 3).

Format of the test. The use of multiple-choice questions (MCQ) in the test did not lead to a significant difference regarding the effect of test preparation, $b = .23$, $se = .13$, 95% CI: −.04 to .49, $p = .094$, indicating that the MCQ format is not a significant moderator of test preparation effect. Similarly, the results of meta-regression (see Table 3) show that the use of short-response items in tests is not a significant moderator, $b = .07$, $se = .16$, 95% CI: −.25 to .39, $p = .683$.

Test administration. The meta-regression indicated that the difference between test administration ways was not significant ($b = -.05$, 95% CI: −.32 to .22, $p = .722$), and the subgroup analysis showed that the effect of test preparation intervention was significant regardless of whether the criterion tests were administered in real testing conditions ($g = .21$, $se = .08$, $k = 27$, $p = .031$) or administered as mocks ($g = .28$, $se = .11$, $k = 66$, $p = .002$).

Test preparation intervention design

Type of test preparation intervention. The subgroup analysis showed that after-school courses ($k = 25$) and in-school courses ($k = 40$) yielded similar significant effects on test performance ($g = .27$, $se = .09$, $p = .055$; and $g = .22$, $se = .13$, $p < .034$, respectively), while courses provided by commercial institutes had a slightly larger effect ($g = .31$, $se = .14$, $k = 19$, $p = .040$). The meta-regression further indicated that when compared to after-school courses, in-school courses and commercial courses did not yield significantly different effects, $b = -.06$, 95% CI: −.39 to .26, $p = .696$; and $b = .01$, 95% CI: −.45 to .48, $p = .954$, respectively. Only two included studies examined the effect of students' self-study, and the effect was not significant, $g = .38$, $se = .08$, $k = 3$, $p = .263$. This result is not conclusive because the number of studies reporting student's self-study was not sufficient.

Breadth of test preparation intervention. Test preparation effects were moderated by the instructional breadth of test preparation intervention. When interventions targeted the same subject or area of the test, but their focus was not narrowed to the exact tested content, the effect was significant, $g = .33$, $se = .10$, $k = 26$, $p = .002$. Similarly, interventions focused solely on the test-specific content and format or drill on the same tested items also had a significant effect, $g = .32$, $se = .09$, $k = 47$, $p < .001$. In comparison, broad intervention focused on general knowledge or skill areas had the smallest effect, which was not significant ($g = -.04$, $se = .13$, $k = 14$, $p = .761$). The meta-regression showed that when compared to broad interventions, specific interventions had significantly larger effects, $b = .39$, 95% CI: .07 to .71, $p = .016$, so did narrow interventions, $b = .37$, 95% CI: .10 to .64, $p = .008$.

Strategy use. The results of the subgroup analysis of the effect of test preparation strategies used during the interventions on student test performance can be found in Table 2. The meta-regression results suggested that test preparation had a significantly larger effect when students were taught test-taking skills during the interventions ($g = .35$, $se = .09$, $k = 62$, $p < .001$) than those that did not ($g = .05$, $se = .12$, $k = 25$, $p = .662$), $b = .31$, 95% CI: .06–.56, $p = .017$. Similarly, interventions that included socio-affective strategies also had a significantly greater effect on students' test performance ($g = .41$, $se = .11$, $k = 24$, $p < .001$) compared with those that did not ($g = .20$, $se = .08$, $k = 63$, $p = .012$), $b = .21$, 95% CI: −.00–.42, $p = .047$. The use of other strategies, such as memorization ($g = .28$, $se = .08$, $k = 59$, $p < .001$) and test preparation management ($g = .30$, $se = .10$, $k = 44$, $p = .003$) also had significant impact on students' test performance according to the subgroup analysis, but when compared to interventions that did not report relevant strategy use, no significant differences were found (see Table 3 for meta-regression results). The effect of using the broad knowledge development strategy was not significant, $g = .17$, $se = .09$, $k = 34$, $p = .069$.

Material use. The moderator analysis also examined whether different material usage in test preparation intervention resulted in the variation of effect sizes. Interventions where students used commercial workbooks had a larger effect on the improvement of test performance ($g = .40$, $se = .10$, $k = 43$, $p < .001$) than those that did not ($g = .15$, $se = .08$, $k = 44$, $p = .077$), and the difference was significant according to the meta-regression ($b = .24$, 95% CI: .01–.47, $p = .044$).

When test preparation interventions provided students with a timed practice test that simulated the real test condition, the effect ($g = .19$, $se = .13$, $k = 27$, $p = .233$) was not significantly different than those that did not ($g = .31$, $se = .08$, $k = 60$, $p < .001$), according to the meta-regression results: $b = -.12$, 95% CI: −.43–.19, $p = .444$. In addition, there was no difference between the interventions that involved sample items to practice ($g = .23$, $se = .08$, $k = 71$, $p < .001$) and those that did not ($g = .10$, $se = .13$, $k = 16$, $p = .794$), $b = .16$, 95% CI: .02–.41, $p = .069$.

Besides, two included studies reported three effect sizes about the effect of test preparation interventions delivered virtually by computer-based programs. The estimate effect was $g = .47$, $se = .32$, $k = 2$, $p = .149$. The meta-regression showed that there was no statistically significant difference between the effect of test

preparation interventions delivered in person (g=.25, se=.08, k=84, p=.001) and those delivered virtually, b=.20, 95% CI: −.46 to .86, p=.546.

Time and duration of the intervention. The meta-regression results showed that the number of contact hours was not a significant moderator of the effect, b=−.006, 95% CI: −.002 to .009, p=.489, nor was the duration of the intervention, b=.01, 95% CI: −.01 to .03, p=.314.

Student-related characteristics

Education level. The test preparation effect was similar for tertiary students (g=.25, se=.13, k=29, p=.046), secondary students (g=.27, se=.11, k=43, p=.019), and primary students (g=.23, se=.19, k=15, p=.220). The differences were not significant, b=−.08, 95% CI: −.65 to .49, p=.777; and b=−.06, 95% CI: −.60 to .48, p=.829.

School type. The subgroup analysis showed that the effect of test preparation programs provided in private schools was significant (g=.43, se=.18, k=8, p=.021), while the effect of programs provided to public school students was not (g=.22, se=.13, k=31, p=.104). The meta-regression did not reveal any significant difference between programs offered by the two types of schools, b=−.21, 95% CI: −.53 to .11, p=.195.

Previous test experience. Among all included studies, 74% of cases reported whether participants had previous testing experience or test preparation experience on the specific test. The subgroup analysis showed that students with previous exposure to the test or test preparation seemed to benefit more from test preparation (g=.69, se=.19, k=6, p<.001); whereas for students without any previous testing experience, the test preparation effect was smaller (g=.06, se=.08, k=58, p=.477). The difference between the two groups of students was significant according to the meta-regression, b=.66, 95% CI: .27–1.05, p=.001, suggesting that students' previous test experience is a significant moderator for test preparation effects.

Study design characteristics

About 53% of included interventions used a randomized design. However, no statistically significant difference of the estimated effect sizes for studies with a random allocation and those without was identified in the meta-regression (b=.20, 95% CI: −.06 to .46, p=.135). In most cases (75%), the test preparation effect was determined by comparing the test performance of the test preparation group (the experimental group) with the results of the control group that did not participate in test preparation. These interventions yielded a significant effect, g=.30, se=.08, k=65, p<.001. In the other cases (25%), the experimental and control conditions took different forms of test preparation, for example, intensive language test preparation courses vs. communicative language courses (for example, Nishitani, 2006; Pan, 2016). These interventions had a smaller and not significant effect: g=.14, se=.12, k=22, p=.224. The meta-regression results showed that

the comparison condition was not a significant moderator of the test preparation effect, $b = .17$, 95% CI: −.09 to .43, $p = .205$.

Sensitivity analysis

To assess the robustness of the findings in this study, we performed a sensitivity analysis. Initially, all studies were included in the meta-regression model. To control for potential bias, we subsequently conducted a follow-up analysis, including only studies with a low risk of bias as assessed by the RoB tools ($n = 25$, $K = 78$). In these studies, a positive overall effect of test preparation was observed ($g = .26$, $SE = .09$, 95% CI = .08–.45, $p < .001$), which was consistent with the effect observed in the full model. The heterogeneity of effect sizes remained significant ($I^2 = 82.37\%$, $Q(77) = 280.32$, $p < .001$), with an estimated variance of true effect sizes of $\tau^2 = .16$.

When the meta-regression models were re-run using only the low-risk-of-bias studies, the significance and direction of most regression coefficients remained unchanged (see Appendix A for details). However, the use of socio-affective strategies, which was a significant moderator in the full model, became nonsignificant ($b = .17$, $SE = .12$, $p = .183$, 95% CI = −.08–.41). While the direction of this effect remained positive, the loss of statistical significance in the low-risk-of-bias model suggests that the evidence for its influence is less robust.

Publication bias

According to the analysis, the publication type, publication year, and the sample size were not significant moderators, indicating that there was less risk of publication bias. There were no significant differences between studies in dissertations or research reports and those that were published in journals (see Table 3 for the meta-regression results). Besides, the risk of publication bias was also evaluated by inspecting the funnel plot (see Figure 5), which gives an impression of the relationship between observed effects and standard error for asymmetry. For this meta-analysis, no systematic relationship was indicated by the funnel plot. The Egger's regression test ($z = -.69$, $p = .501$) also suggested the risk of publication bias was judged to be low.

Discussion

The results of this meta-analysis show that test preparation has a statistically significant overall effect on test performance across a broad range of different domains and different tests in authentic educational contexts. Our findings suggest that the test preparation effect is robust across different test conditions, for both large-scale high-stakes and low-stakes tests, curriculum-based and non-curriculum-based tests.

The test preparation effect can be explained by four groups of explanations, as elaborated by Arendasy et al. (2016) and Lievens et al. (2007). First, test preparation may reduce construct-irrelevant factors, such as unfamiliarity and test anxiety, leading to an increase in test scores. Second, the development of test-wiseness and test-taking skills during test preparation increases the likelihood of solving test items correctly. Third, domain-specific knowledge or test-specific cognitive

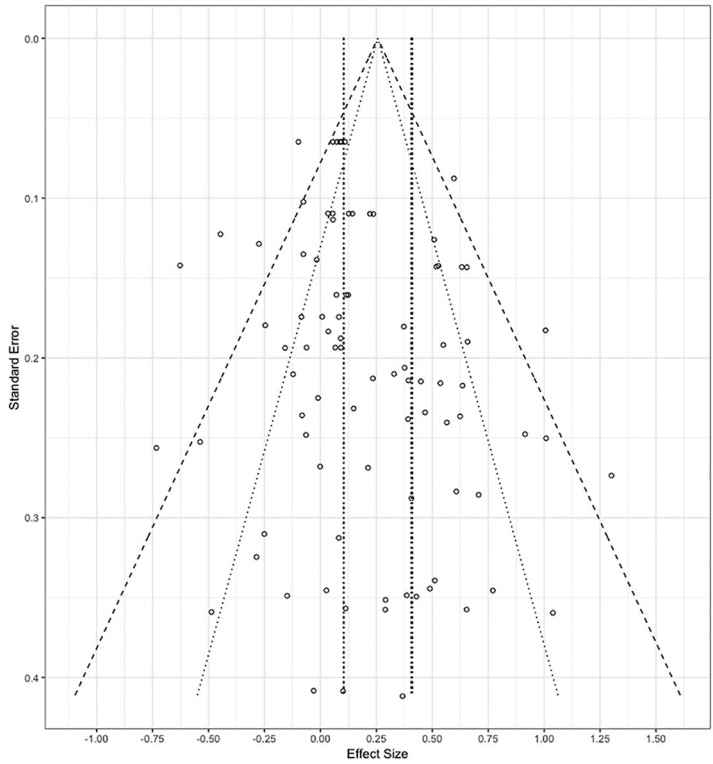

FIGURE 5. *A funnel plot showing the relationship between standard error and observed effect size for the meta-analysis.*

skills could be developed by test preparation. Fourth, generalizable latent cognitive ability and/or broad knowledge could be enhanced by test preparation intervention.

The moderator analysis of the observed test preparation effects provides some support for the first three explanations, as discussed in the following. However, the fourth model was not supported by our meta-analysis. Our finding suggests that the test preparation effect is unlikely to transfer to other tests, which is in line with previous studies (see Arendasy et al., 2016; Freund & Holling, 2011; Koenig et al., 2008) that found that preparing for a test by either coaching or repeated practice on sample items failed to improve general cognitive ability.

The first and second models of test preparation effects were corroborated by this meta-analysis. The analysis shows that students who had some previous test or test preparation experience possibly benefited from test preparation more than those who had not. This indicates how increasing test familiarity can help students prepare for tests more effectively. We also found that test preparation interventions that included instructions or practices on test-taking skills had a significantly

greater effect than those that did not include test-taking skills training. However, test-taking skills such as guessing, wrong answer elimination, and time management may not be able to influence the underlying construct the test examined. Such score increases might not represent improvements in the underlying construct highlighted by the test, thus undermining the validity of the interpretation of test scores.

In this meta-analysis, the breadth of test preparation intervention was categorized as narrow, specific, and broad, differentiating the alignment of teaching/learning content and the test. Findings indicate that interventions with a specific test-related focus and ones that even solely drilled students on tested content had a significantly larger effect on test performance than those that developed generalizable skills. Although test-oriented teaching and learning has received a lot of criticism, our findings confirm its effectiveness in increasing test scores and probably provides some empirical support for the first and third model suggested by Arendasy et al. (2016), according to which the increases of test scores could be attributed to test familiarization and an increase of domain-specific knowledge or specific cognitive skills (Xie, 2013).

In fact, the framework of models of test preparation effect was mainly built upon evidence of retest effect on cognitive ability tests (Lievens et al., 2007). To some extent, the retest effect and test preparation effect shared similarities in terms of mechanisms. For example, both retesting and test preparation allow students to familiarize themselves with the test and test-taking skills. Additionally, retesting is an important and widely utilized way of test preparation. Therefore, it is plausible to explain the effects of test preparation using the framework of causes of retest effects. However, applying this framework directly to test preparation contexts has unavoidable limits due to the natural differences between retesting and test preparation. Compared to retesting, test preparation is a more complicated learning process that includes various cognitive and metacognitive components, where students interact dynamically with not only test papers and relevant materials but also their teachers, peers, and even their parents.

This meta-analysis identifies that using workbooks is a significant moderator of the test preparation effect, as is developing test preparation management strategies during test preparation intervention; both significantly benefited test performance. These findings cannot be explained by the current framework drawn from retest effects (Arendasy et al., 2016; Lievens et al., 2007). Instead, learning and cognition theories provide hints as to how to explain the findings.

A test workbook typically contains an introduction to the test structure, test format, sample test questions and answers, and techniques, as well as some practice tests and review of knowledge content (see *Princeton Review SAT Prep*, for example), which provides scaffolding to help students build the schematic structure of the test and test content, which can further guide test preparation. A large number of studies confirmed the positive effect of scaffolding on educational achievements (Jitendra et al., 2009; Wijekumar et al., 2014). Nevertheless, we should be aware that scaffolding usually refers to "a general teaching strategy and concept where educators provide guidance or other forms of support such as

prompts, cues, instructions, information organization, or reference materials" (Perry et al., 2021, p. 104), and previous studies primarily investigated the effect of scaffolding by manipulating classroom instructions, with no existing studies focused on the effect of scaffolding by using workbooks. More research is needed to support or reject this interpretation of the test preparation effect moderated by workbooks.

The meta-regression analysis of all included studies demonstrated that incorporating socio-affective strategies into the test preparation process was a significant moderator of test preparation effects. A possible explanation is that working collaboratively with teachers and peers allows students to share information about the test and gain feedback about their performance, which further optimizes their learning. Literature on collaborative learning showed that students can benefit from collaborative work, specifically for solving high-complexity tasks with more cognitive load (see Kirschner et al., 2011; Retnowati et al., 2017). Besides, some studies suggested that the use of socio-affective strategies can ease students' test anxiety (Saeidi & Khaliliaqdam, 2013), leading to an increase in test scores. However, in the meta-regression restricted to studies with low risk of bias, the effect of socio-affective strategies on test preparation did not reach statistical significance. This reduction in significance indicates that the influence of socio-affective strategies on test preparation effects might be exaggerated due to potential methodological flaws or biases. Future research with higher-quality design may provide more definitive conclusions about the role of socio-affective strategies in improving test preparation outcomes.

An unexpected but important finding of this meta-analysis is that sample items and timed practice tests were not significant moderators of the effect of test preparation when analyzed across studies. In previous meta-analyses of practice effect on cognitive ability tests (Hausknecht et al., 2007; Kulik et al., 1984; Scharfen et al., 2018), the use of sample items and practice tests has shown great impact on memory retention and the improvement of cognitive abilities. However, this meta-analysis finds little evidence of practice effect. The reasons might be that compared to cognitive ability tests, educational tests assess not only cognitive abilities, but also the understanding, evaluation, and reflection of knowledge, as well as problem-solving ability and critical thinking ability. This might also be the reason why memorization strategy is not a significant moderator in our study. Preparing for educational tests should be a comprehensive process that covers aspects of different skills other than knowledge retention. Besides, retest effects were only robust when using identical test forms in retesting sessions, while the retest effect for alternative forms of tests was found to be significantly smaller (Scharfen et al., 2018). In real educational contexts, it is rarely possible to practice test papers or items that are identical to those in real tests. We therefore infer that the practice effect can be smaller in educational tests. More quantitative and experimental studies are needed to test this inference.

Another unexpected finding of this study is that the effects of test preparation were significant for both high-stakes and low-stakes tests, but there was no difference between the two types of tests. This result seems to be unreasonable, as

previous studies have shown that students were more motivated and tried harder in high-stakes tests (Knekta & Sundström, 2019). One of the possible explanations for this finding is that the results were derived from (quasi) experimental studies in which participants were given a test preparation intervention and were required to prepare for the test regardless of the test's stakes. Such a design did not simulate real-world situations, where students might not study or pay less effort for low-stakes tests. Another reason could be the self-selection bias: students who opted to participate in the (quasi) experiments might be highly motivated and better engaged than those who did not participate. Thus, whether the test was high-stakes or low-stakes did not affect students' learning and thus did not moderate test preparation effects.

According to this meta-analysis, the effect of commercial test preparation courses did not significantly exaggerate additional test preparation courses provided by schools or universities. Besides, there was no significant difference in effectiveness between courses provided by private schools and public schools. Actually, there has been a long debate on educational equity issues and commercial educational courses (shadow education) in previous literature (Bray, 1999; Zhang & Bray, 2020). Some researchers were concerned that students from low SES groups might be disadvantaged because they cannot get equal access to commercial courses (Buchmann et al., 2010; Zhang, 2013). However, it is not clear from the current state of the literature that there is good evidence for how the quality of test preparation instruction varies, or whether test preparation activities are more prevalent in certain kinds of schools. More evidence from future research is needed.

Lastly, compared to test preparation courses (coaching) provided by teachers, students' self-test preparation was under-investigated. Current experimental studies examined whether a specific instruction or teaching pattern was effective, regarding test preparation as teacher-driven activities, while students' potentials and initiatives were neglected. With more studies and better theory, it will be possible to identify features and dimensions of students' test preparation practices that are responsible for greater score gains.

Limitations

There were some limitations caused by the characteristics of some of the primary studies included in the meta-analysis. First, many studies provided insufficient information on the test preparation intervention, making it difficult to code some moderators. For example, most studies described the procedures of test preparation intervention very briefly without detailed description of teachers' instructions. Therefore, we were unable to code the quality of instructions and did not include this factor in the moderator analysis. Second, some included studies that evaluated the effect of a long-time test preparation intervention—that is, a course that took 2 hours per week and lasted for 30 weeks, though they used a random allocation for participants. It can be problematic to attribute the score changes simply to the intervention without appropriate techniques for controlling other factors (such as other activities related to test preparation), thus the effect size might be inflated.

Besides, the findings in the meta-analysis are primarily based on studies carried out on college admission tests in the U.S. context. Although the findings indicate that the magnitude of the effects in the United States was similar in other contexts, and the difference between different tests was not significant, some caution should be taken when generalizing results to other testing contexts, especially tests of other domains (i.e., science, medicine, foreign language except English) or in less investigated regions (i.e., regions in the Middle East, Southeast Asia, and Africa).

There were also limitations in terms of the interpretation of the results. There was no solid theory developed in the test preparation area to analyze its process and explain the effect, thus we interpreted the test preparation effect mainly based on the framework of retesting and some theories from cognitive science, which might not fit well and leave out of the scope other possible frameworks, such as self-regulation or instructional support. However, it is difficult to construct a valid framework for test preparation effect based on those studies without a rigorous design or when lacking precise descriptions of teaching and learning strategies used for test preparation. More rigorous and thoroughly described experiments on specific material use and strategy use during test preparation are needed.

Conclusion

This meta-analysis provides evidence of a significant positive effect of test preparation on test performance, particularly those test preparation interventions focused on relevant content and skills or drilled on tested content. It also highlights the moderating effects of the use of workbooks, encouraging the use of socio-affective strategies and the development of test-taking skills. This study serves as a comprehensive quantitative synthesis of test preparation, contributing to the growing literature on the effectiveness of test preparation and advancing our understanding of its strengths and limitations. Further research on test preparation should include greater details of teachers' instructions and students' practices during the test preparation process, thus providing additional insights for using test preparation as effective learning and teaching activities.

Appendix A

TABLE A1

Results of meta-regressions of studies with low risk of bias

	b	*SE*	95% CI Low	95% CI Upper	*p*
Nature of the Test					
Test Stakes F (df1 = 2, df2 = 75) = 1.72, *p* = .186					
Intercept	.40	.11	.18	.63	< .001
Low-stakes tests	−.22	.14	−.51	.06	.121
Sample size	−.00	.00	−.00	.00	.248
Curriculum F (df1 = 2, df2 = 75) = .57, *p* = .570					
Intercept	.35	.12	.11	.58	.004
Curriculum-based tests	−.11	.18	−.46	.25	.549
Sample size	−.00	.00	−.00	.00	.505
Type of tests F (df1 = 2, df2 = 75) = 1.96, *p* = .148					
Intercept	.46	.13	.20	.71	< .001
College admission tests	−.27	.16	−.59	.04	.088
Sample size	−.00	.00	−.00	.00	.491
Subject Domain F (df1 = 5, df2 = 72) = 0.48, *p* = 0.787					
Intercept	.34	.13	.09	.60	.009
Verbal	−.07	.13	−.33	.19	.606
Mathematics	−.01	.13	−.27	.26	.966
Others	−.11	.14	−.40	.18	.435
Mixed	−.01	.16	−.32	.31	.960
Sample size	−.00	.00	−.00	.00	.384
Test format					
Included MCQ Questions F (df1 = 2, df2 = 75) = 2.02, *p* = 0.139					
Intercept	.12	.15	−.18	.43	.412
Yes	.25	.14	−.03	.54	.083
Sample size	−.00	.00	−.00	.00	.286
Included SR Questions F (df1 = 2, df2 = 75) = 0.77, *p* = 0.467					
Intercept	.31	.11	.10	.53	.004
Yes	.13	.16	−.19	.46	.406
Sample size	−.00	.00	−.00	.00	.279
Test Administration F (df1 = 2, df2 = 75) = 0.43, *p* = 0.650					
Intercept	.33	.11	.10	.55	.005
Real conditions	−.04	.14	−.31	.24	.800
Sample size	−.00	.00	−.00	.00	.417
Type of Test Preparation Intervention F (df1 = 4, df2 = 73) = 0.35, *p* = 0.846					
Intercept	.37	.18	.01	.72	*.044*
In-school courses	−.10	.17	−.43	.23	.558
Commercial	.06	.25	−.44	.57	.809
Self-study	−.06	.56	−1.17	1.04	.909
Sample size	−.00	.00	−.00	.00	.379

(continued)

TABLE A1 (CONTINUED)

	b	*SE*	95% CI Low	95% CI Upper	*p*
Level of Test Preparation Intervention F (df1 = 3, df2 = 74) = 1.94, *p* = 0.107					
Intercept	.05	.19	−.33	.43	.802
Specific	.33	.20	.07	.72	.043
Narrow	.37	.21	.09	.75	.041
Sample size	−.00	.00	−.00	.00	.334
Strategy Use for Test Preparation					
Test Preparation Management F (df1 = 2, df2 = 75) = 0.39, *p* = 0.680					
Intercept	.32	.16	.00	.64	.047
Yes	−.01	.17	−.34	.33	.973
Sample size	−.00	.00	−.00	.00	.393
Test-Taking Skills F (df1 = 2, df2 = 75) = 4.97, *p* = 0.086					
Intercept	.09	.19	−.18	.46	.633
Yes	.27	.17	.08	.60	.036
Sample size	−.00	.00	−.00	.00	.703
Memorization F (df1 = 2, df2 = 75) = 0.45, *p* = 0.639					
Intercept	.33	.12	.09	.58	.008
Yes	−.02	.09	−.20	.16	.841
Sample size	−.00	.00	−.00	.00	.357
Socio-affective Strategies F (df1 = 2, df2 = 75) = 1.43, *p* = .247					
Intercept	.28	.10	.07	.49	.009
Yes	.17	.12	−.08	.41	.183
Sample size	−.00	.00	−.00	.00	.257
Broad Knowledge/Skill Development F (df1 = 2 df2 = 75) = .56, *p* = .576					
Intercept	.34	.11	.12	.56	.003
Yes	−.05	.10	−.25	.15	.594
Sample size	−.00	.00	−.00	.00	.391
Material Use for Test Preparation					
Workbook F (df1 = 2, df2 = 75) = 1.99, *p* = .143					
Intercept	.25	.13	−.02	.51	.067
Yes	.22	.14	.01	.42	.047
Sample size	−.00	.00	−.00	.00	.439
Practice Tests F (df1 = 2, df2 = 75) = .77, *p* = .467					
Intercept	.38	.13	.13	.63	.004
Yes	−.13	.16	−.46	.19	.419
Sample size	−.00	.00	−.00	.00	.283
Sample Items F (df1 = 2, df2 = 75) = .99, *p* = .375					
Intercept	.17	.17	−.17	.52	.324
Yes	.17	.16	−.15	.48	.291
Sample size	−.00	.00	−.00	.00	.335
Computer Programs F (df1 = 2, df2 = 75) = .52, *p* = .595					
Intercept	.32	.11	.10	.53	.004
Yes	.22	.48	−.73	1.18	.642
Sample size	−.00	.00	−.00	.00	.379

(continued)

TABLE A1 (CONTINUED)

	b	*SE*	95% CI Low	95% CI Upper	*p*
Time Use for Test Preparation					
Contact Hours F (df1 = 2, df2 = 75) = .60, *p* = .549					
Intercept	.33	.11	.12	.54	.003
Contact hours	−.00	.00	−.00	.00	.575
Sample size	−.00	.00	−.00	.00	.405
Duration F (df1 = 2, df2 = 75) = .93, *p* = .399					
Intercept	.22	.14	−.06	.51	.120
Duration	.01	.01	−.01	.03	.296
Sample size	−.00	.00	−.00	.00	.397
Students' Education Level F (df1 = 3, df2 = 74) = .26, *p* = .854					
Intercept	.40	.32	−.25	1.04	.225
Secondary	−.08	.30	−.69	.53	.796
Tertiary	−.07	.32	−.70	.56	.825
Sample size	−.00	.00	−.00	.00	.395
School Type F (df1 = 5, df2 = 72) = .52, *p* = .764					
Intercept	.46	.20	.07	.86	.023
Public school	−.21	.16	−.53	.10	.185
Commercial institute	−.07	.31	−.68	.54	.822
University	−.17	.24	−.66	.32	.467
Others	−.12	.32	−.76	.51	.696
Sample size	−.00	.00	−.00	.00	.496
Students' Previous Test or Test Preparation Experience F (df1 = 3, df2 = 74) = 7.45, *p* < .001					
Intercept	.14	.11	−.09	.36	.227
Yes	.64	.25	.14	1.14	.012
Not mention	.49	.12	.25	.72	< .001
Sample size	−.00	.00	−.00	.00	.395
Random Allocation F (df1 = 2, df2 = 75) = 1.15, *p* = .321					
Intercept	.22	.13	−.04	.49	.100
Yes	.17	.15	−.12	.45	.253
Sample size	−.00	.00	−.00	.00	.460
Comparison F (df1 = 2, df2 = 75) = 1.36, *p* = .264					
Intercept	.20	.14	−.07	.47	.137
No test preparation	.18	.13	−.08	.45	.175
Sample size	−.00	.00	−.00	.00	.312
Publication Type F (df1 = 3 df2 = 74) = 1.43, *p* = .241					
Intercept	.39	.17	.04	.73	.030
Journal	−.10	.19	−.48	.29	.624
Others	.39	.31	−.22	1.00	.207
Sample size	−.00	.00	−.00	.00	.127
Publication Year F (df1 = 2, df2 = 75) = .38, *p* = .682					
Intercept	−.17	14.05	−28.15	27.82	.991
Year	.00	.00	−.01	.01	.973
Sample size	−.00	.00	−.00	.00	.395
Sample Size F (df1 = 1, df2 = 76) = .84, *p* = .362					
Intercept	.32	.11	.11	.53	.003
Year	−.00	.00	−.00	.00	.362

ORCID iDs

Zhanxin Hao https://orcid.org/0000-0002-1526-8515
Kit Double https://orcid.org/0000-0001-8120-1573

References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Alderman, D. L., & Powers, D. E. (1980). The effects of special preparation on SAT-verbal scores. *American Educational Research Journal*, *17*(2), 239–251. https://doi.org/10.3102/00028312017002239

Allalouf, A., & Ben-Shakhar, G. (1998). The effect of coaching on the predictive validity of scholastic aptitude tests. *Journal of Educational Measurement*, *35*(1), 31–47. https://doi.org/10.1111/j.1745-3984.1998.tb00526.x

Anastasi, A. (1981). Coaching, test sophistication, and developed abilities. *American Psychologist*, *36*, 1086–1093. https://doi.org/10.1037/0003-066X.36.10.1086

Appelrouth, J. I., Zabrucky, K. M., & Moore, D. (2017). Preparing students for college admissions tests. *Assessment in Education: Principles, Policy & Practice*, *24*(1), 78–95. https://doi.org/10.1080/0969594X.2015.1075958

Arendasy, M. E., Sommer, M., Gutiérrez-Lobos, K., & Punter, J. F. (2016). Do individual differences in test preparation compromise the measurement fairness of admission tests? *Intelligence*, *55*, 44–56. https://doi.org/10.1016/j.intell.2016.01.004

Bailey, D., & Christopher, J. (2018). The effects of online collaborative writing and TOEIC writing test-preparation on L2 writing performance. *The Journal of AsiaTEFL*, *15*, 383–397. https://doi.org/10.18823/asiatefl.2018.15.2.8.383

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1983). Effects of coaching programs on achievement test performance. *Review of Educational Research*, *53*(4), 571–585. https://doi.org/10.2307/1170221

Banks, G. C., Kepes, S., & Banks, K. P. (2012). Publication bias: The antagonist of meta-analytic reviews and effective policymaking. *Educational Evaluation and Policy Analysis*, *34*(3), 259–277. https://doi.org/10.3102/0162373712446144

Becker, B. J. (1990). Coaching for the scholastic aptitude test: Further synthesis and appraisal. *Review of Educational Research*, *60*(3), 373–417. https://doi.org/10.3102/00346543060003373

Bol, L., & Hacker, D. J. (2001). A comparison of the effects of practice tests and traditional review on performance and calibration. *The Journal of Experimental Education*, *69*(2), 133–151. https://doi.org/10.1080/00220970109600653

Bookman, A. B. (1981). *The effects of preparation for tests on standardized mathematics achievement test growth*. University of Connecticut.

Borenstein, M., Cooper, H., Hedges, L., & Valentine, J. (2009). Effect sizes for continuous data. *The handbook of research synthesis and meta-analysis*, *2*, 221–235.

Bray, M. (1999). *The Shadow education system: Private tutoring and its implications for planners*. UNESCO, International Institute for Educational Planning.

Briggs, D. C. (2001). The effect of admissions test preparation: Evidence from NELS:88. *Chance*, *14*(1), 10–18. https://doi.org/10.1080/09332480.2001.10542245

Briggs, D. C. (2009). *Preparation for college admission exams* (2009 NACAC Discussion Paper). National Association for College Admission Counseling. https://eric.ed.gov/?id=ED505529

Brunner, M., Artelt, C., Krauss, S., & Baumert, J. (2007). Coaching for the PISA test. *Learning and Instruction*, *17*(2), 111–122. https://doi.org/10.1016/j.learninstruc.2007.01.002

Buchmann, C., Condron, D. J., & Roscigno, V. J. (2010). Shadow education, American style: Test preparation, the SAT and college enrollment. *Social Forces*, *89*(2), 435–461. https://doi.org/10.1353/sof.2010.0105

Bunting, B. P., & Mooney, E. (2001). The effects of practice and coaching on test results for educational selection at eleven years of age. *Educational Psychology*, *21*(3), 243–253. https://doi.org/10.1080/01443410120065450

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*(2), 115–144. https://doi.org/10.1177/2515245919847196

Cheung, A., & Slavin, R. E. (2016). How methodological features of research studies affect effect sizes. *Educational Researcher*, *45*(5), 283–292. https://doi.org/10.3102/0013189x16656615

Domingue, B., & Briggs, D. (2009). Using linear regression and propensity score matching to estimate the effect of coaching on the SAT. *General Linear Model Journal*, *35*(1), 12–29.

Farnsworth, T. (2013). Effects of targeted test preparation on scores of two tests of oral English as a second language. *TESOL Quarterly*, *47*(1), 148–156. https://www.jstor.org/stable/43267777

Filizola, E. (2008). *The effect of a test preparation course on the SAT scores of students at Saint Joseph Academy*. University of Houston.

Freund, P. A., & Holling, H. (2011). How to get really smart: Modeling retest and training effects in ability testing using computer-generated figural matrix items. *Intelligence*, *39*(4), 233–243. https://doi.org/10.1016/j.intell.2011.02.009

Gaye, Z. Z. (2001). *The ACT Prep and its effects on students who participated in the Multicultural Educational Enrichment (MEE) program at Tennessee State University*. Tennessee State University.

Green, A. (2007). Washback to learning outcomes: A comparative study of IELTS preparation and university pre-sessional language courses. *Assessment in Education: Principles, Policy & Practice*, *14*(1), 75–97. https://doi.org/10.1080/09695940701272880

Griffin, B., Carless, S., & Wilson, I. (2013). The effect of commercial coaching on selection test performance. *Medical Teacher*, *35*(4), 295–300. https://doi.org/10.3109/0142159X.2012.746451

Hardison, C. M., & Sackett, P. R. (2008). Use of writing samples on standardized tests: Susceptibility to rule-based coaching and the resulting effects on score improvement. *Applied Measurement in Education*, *21*(3), 227–252. https://doi.org/10.1080/08957340802161782

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge. https://doi.org/10.4324/9780203887332

Hausknecht, J., Halpert, J., Paolo, N., & Gerrard, M. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology*, *92*, 373–385. https://doi.org/10.1037/0021-9010.92.2.373

Hensley, L., Kulesza, A., Peri, J., Brady, A. C., Wolters, C. A., Sovic, D., & Breitenberger, C. (2021). Supporting undergraduate biology students' academic success: Comparing

two workshop interventions. *CBE—Life Sciences Education*, *20*(4), ar60. https://doi.org/10.1187/cbe.21-03-0068

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177. https://doi.org/10.1111/j.1750-8606.2008.00061.x

Holmes, C. T., & Keffer, R. L. (1995). A computerized method to teach Latin and Greek root words: Effect on verbal SAT scores. *The Journal of Educational Research*, *89*(1), 47–50. https://doi.org/10.1080/00220671.1995.9941192

Holthaus, R. A. (2008). *The effects of a high school coaching program on SAT-Mathematics test scores*. Walden University.

Hong, E., & Peng, Y. (2008). Do Chinese students' perceptions of test value affect test performance? Mediating role of motivational and metacognitive regulation in test preparation. *Learning and Instruction*, *18*(6), 499–512. https://doi.org/10.1016/j.learninstruc.2007.10.002

Hsiao, T. Y., & Oxford, R. L. (2002). Comparing theories of language learning strategies: A confirmatory factor analysis. *The modern language journal*, *86*(3), 368–383.

Hu, R., & Trenkic, D. (2021). The effects of coaching and repeated test-taking on Chinese candidates' IELTS scores, their English proficiency, and subsequent academic achievement. *International Journal of Bilingual Education and Bilingualism*, *24*(10), 1486–1501. https://doi.org/10.1080/13670050.2019.1691498

Jitendra, A. K., Star, J. R., Starosta, K., Leh, J. M., Sood, S., Caskie, G., Hughes, C. L., & Mack, T. R. (2009). Improving seventh grade students' learning of ratio and proportion: The role of schema-based instruction. *Contemporary Educational Psychology*, *34*(3), 250–264. https://doi.org/10.1016/j.cedpsych.2009.06.001

Justus, L. C. (2010). *Impact of a school-based test preparation course on ACT scores with consideration of cultural associations*. Union University.

Kirschner, F., Paas, F., & Kirschner, P. A. (2011). Task complexity as a driver for collaborative learning efficiency: The collective working-memory effect. *Applied Cognitive Psychology*, *25*(4), 615–624. https://doi.org/10.1002/acp.1730

Kitsantas, A. (2002). Test preparation and performance: A self-regulatory analysis. *The Journal of Experimental Education*, *70*(2), 101–113. https://doi.org/10.1080/00220970209599501

Knekta, E., & Sundström, A. (2019). It was, perhaps, the most important one': Students' perceptions of national tests in terms of test-taking motivation. *Assessment in Education: Principles, Policy & Practice*, *26*(2), 202–221. http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-131024

Koenig, K. A., Frey, M. C., & Detterman, D. K. (2008). ACT and general cognitive ability. *Intelligence*, *36*(2), 153–160. https://doi.org/10.1016/j.intell.2007.03.005

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Kulik, J. A., Bangert-Drowns, R. L., & Kulik, C. C. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin*, *95*(2), 179–188. https://doi.org/10.1037/0033-2909.95.2.179

Kulik, J. A., Kulik, C.-L. C., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal*, *21*(2), 435–447. https://doi.org/10.3102/00028312021002435

Lai, E. R., & Waltman, K. (2008). Test preparation: Examining teacher perceptions and practices. *Educational Measurement: Issues and Practice*, *27*(2), 28–45. https://doi.org/10.1111/j.1745-3992.2008.00120.x

Lane, K. L., Kalberg, J. R., Mofield, E., Wehby, J. H., & Parks, R. J. (2009). Preparing students for college entrance exams: Findings of a secondary intervention conducted within a three-tiered model of support. *Remedial and Special Education*, *30*(1), 3–18. https://doi.org/10.1177/0741932507314022

Lee, J.-Y. (2019). Pedagogical effects of teaching test-taking strategies to EFL college students. *Reading in a Foreign Language*, *31*(2), 226–248. https://eric.ed.gov/?id=EJ1232384

Li, H., & Xiong, Y. (2018). The relationship between test preparation and state test performance: Evidence from the Measure of Effective Teaching (MET) project. *Education Policy Analysis Archives*, *26*, 64–64. https://doi.org/10.14507/epaa.26.3530

Lievens, F., Reeve, C. L., & Heggestad, E. D. (2007). An examination of psychometric bias due to retesting on cognitive ability tests in selection settings. *Journal of Applied Psychology*, *92*(6), 1672–1682. https://doi.org/10.1037/0021-9010.92.6.1672.

Lin, T.-C., Hsu, Y.-S., Lin, S.-S., Changlai, M.-L., Yang, K.-Y., & Lai, T.-L. (2012). A Review of empirical evidence on scaffolding for science education. *International Journal of Science and Mathematics Education*, *10*(2), 437–455. https://doi.org/10.1007/s10763-011-9322-z

Madaus, G. F. (1988). The distortion of teaching and testing: High-stakes testing and instruction. *Peabody Journal of Education*, *65*(3), 29–46. https://doi.org/10.1080/01619568809538611

McGaghie, W. C., Downing, S. M., & Kubilius, R. (2004). What is the impact of commercial test preparation courses on medical examination performance? *Teaching and Learning in Medicine*, *16*(2), 202–211. https://doi.org/10.1207/s15328015tlm1602_14

McMann, P. K. (1994). *The effects of teaching practice review items and test-taking strategies on the ACT mathematics scores of second-year algebra students*. Wayne State University.

Messick, S. (1982). Issues of effectiveness and equity in the coaching controversy: Implications for educational and testing practice. *Educational Psychologist*, *17*(2), 67–91. https://doi.org/10.1080/00461528209529246

Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin*, *89*(2), 191–216. https://doi.org/10.1037/0033-2909.89.2.191

Mitchell, I., Nistor, N., Baltes, B., & Brown, M. (2016). Effect of vocabulary test preparation on low-income Black middle school students' reading scores. *Journal of Educational Research and Practice*, *6*(1), 105–118. https://eric.ed.gov/?id=EJ1132300

Montgomery, P., & Lilly, J. (2012). Systematic reviews of the effects of preparatory courses on university entrance examinations in high school-age students. *International Journal of Social Welfare*, *21*(1), 3–12. https://doi.org/10.1111/j.1468-2397.2011.00812.x

Moore, R., Sanchez, E., & San Pedro, M. O. (2018). *Investigating test prep impact on score gains using quasi-experimental propensity score matching* (ACT Working Paper No. 2018-6). ACT, Inc. https://eric.ed.gov/?id=eD593130

Moss, G. L., Chippendale, E. K., Mershon, C. W., & Carney, T. (2012). Effects of a coaching class on the ACT scores of students at a Large Midwest High School. *Journal of College Admission*, *217*, 16–23.

Nishitani, A. (2006). Teaching grammar for the TOEIC test: Is test preparation instruction effective? *Glottodidactica*, *32*, 139–146. http://hdl.handle.net/10593/2298

Oxford, R. L. (1991). *Language learning strategies: What every teacher should know*. Boston, MA: Heinle & Heinle Publishers.

Pan, Y.-C. (2016). Traditional and non-traditional test preparation practices: Learner performance and perspectives. *Electronic Journal of Foreign Language Teaching*, *13*(2), 170–183.

Parrott, T. N. (2012). *ACT test preparation course and its impact on students' college- and career-readiness*. Tennessee State University.

Perry, T., Lea, R., Jørgensen, C. R., Cordingley, P., Shapiro, K., & Youdell, D. (2021). *Cognitive science in the classroom*. Education Endowment Foundation (EEF). https://educationendowmentfoundation.org.uk/evidence-summaries/evidence-reviews/cognitive-science-approaches-in-the-classroom/

Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. *Language testing*, *20*(1), 26–56.

Powers, D. E. (1986). Relations of test item characteristics to test preparation/test practice effects: A quantitative summary. *Psychological Bulletin*, *100*(1), 67–77. https://doi.org/10.1037/0033-2909.100.1.67

Powers, D. E., & Rock, D. A. (1999). Effects of con SAT I: Reasoning test scores. *Journal of Educational Measurement*, *36*(2), 93–118. https://doi.org/10.1111/j.1745-3984.1999.tb00549.x

Powers, D. E., & Swinton, S. S. (1984). Effects of self-study for coachable test item types. *Journal of Educational Psychology*, *76*(2), 266–278. https://doi.org/10.1037/0022-0663.76.2.266

Purpura, J. E. (1999). *Learner strategy use and performance on language tests* (Vol. 8). Cambridge University Press.

R Core Team. (2019). *R: A language and environment for statistical computing*. R foundation for Statistical Computing.

Rao, C., McPherson, K., Chand, R., & Khan, V. (2003). Assessing the impact of IELTS preparation programs on candidates' performance on the General Training reading and writing test modules. *International English Language Testing System (IELTS) Research Reports*, *5*, 236–262. https://doi.org/10.3316/informit.078648215906459

Retnowati, E., Ayres, P., & Sweller, J. (2017). Can collaborative learning improve the effectiveness of worked examples in learning mathematics? *Journal of Educational Psychology*, *109*(5), 666. https://doi.org/10.1037/edu0000167

Reynolds, A. J., Oberman, G. L., & Perlman, C. (1988). An analysis of a PSAT coaching program for urban gifted students. *The Journal of Educational Research*, *81*(3), 155–164. https://doi.org/10.1080/00220671.1988.10885816

Robb, T. N., & Ercanbrack, J. (1999). A study of the effect of direct test preparation on the TOEIC scores of Japanese university students. *TESL-EJ*, *3*(4), 1–22.

Ross, S. J. (2008). Language testing in Asia: Evolution, innovation, and policy challenges. *Language Testing*, *25*(1), 5–13. https://doi.org/10.1177/0265532207083741

Saeidi, M., & Khaliliaqdam, S. (2013). The effect of socio-affective strategies on students' test anxiety across different genders. *Theory & Practice in Language Studies*, *3*(2), 269–274. https://doi.org/10.4304/tpls.3.2.269-274

Scharfen, J., Jansen, K., & Holling, H. (2018). Retest effects in working memory capacity tests: A meta-analysis. *Psychonomic Bulletin & Review*, *25*(6), 2175–2199. https://doi.org/10.3758/s13423-018-1461-6

Schueler, B., & Rodriguez-Segura, D. (2020). *Can camp get you into a good secondary school? A field experiment of targeted instruction in Kenya* (Technical Report No. 197). Annenberg Institute at Brown University.

Slavin, R. E., & Smith, D. (2009). The relationship between sample sizes and effect sizes in systematic review in education. *Educational Evaluation and Policy Analysis*, *31*(4), 500–506. https://doi.org/10.3102/0162373709352369

Steel, P. D., & Kammeyer-Mueller, J. D. (2002). Comparing meta-analytic moderator estimation techniques under realistic conditions. *Journal of Applied Psychology*, *87*(1), 96.

Sterne, J. A. C., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., Henry, D., Altman, D. G., Ansari, M. T., Boutron, I., Carpenter, J. R., Chan, A.-W., Churchill, R., Deeks, J. J., Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y. K., Pigott, T. D., & . . . Higgins, J. P. T. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*, *355*, i4919. https://doi.org/10.1136/bmj.i4919

Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H.-Y., Corbett, M. S., & Eldridge, S. M. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. *BMJ*, *366*, l4898. https://doi.org/10.1136/bmj.l4898

Takallou, F., Vahdany, F., Araghi, S. M., & Tabrizi, A. R. N. (2015). The effect of test taking strategy instruction on Iranian high school students' performance on English section of the university entrance examination and their attitude towards using these strategies. *International Journal of Applied Linguistics and English Literature*, *4*(6), 119–129. https://doi.org/10.7575/aiac.ijalel.v.4n.6p.119

Tanner-Smith, E. E., Tipton, E., & Polanin, J. R. (2016). Handling complex meta-analytic data structures using robust variance estimates: A tutorial in R. *Journal of Developmental and Life-Course Criminology*, *2*(1), 85–112. https://doi.org/10.1007/s40865-016-0026-5

Trenkic, D., & Hu, R. (2021). Teaching to the test: The effects of coaching on English-proficiency scores for university entry. *Journal of the European Second Language Association*, *5*(1), 1–15. https://doi.org/10.22599/jesla.74

Van Aert, R. C., Wicherts, J. M., & Van Assen, M. A. (2019). Publication bias examined in meta-analyses from psychology and medicine: A meta-meta-analysis. *PloS one*, *14*(4), e0215052.

Wijekumar, K., Meyer, B. J. F., Lei, P.-W., Lin, Y.-C., Johnson, L. A., Spielvogel, J. A., Shurmatz, K. M., Ray, M., & Cook, M. (2014). Multisite randomized controlled trial examining intelligent tutoring of structure strategy for fifth-grade readers. *Journal of Research on Educational Effectiveness*, *7*(4), 331–357. https://doi.org/10.1080/19345747.2013.853333

Winke, P., & Lim, H. (2017). The effects of test preparation on second-language listening test performance. *Language Assessment Quarterly*, *14*(4), 380–397. https://doi.org/10.1080/15434303.2017.1399396

Xie, Q. (2013). Does test preparation work? Implications for scores. *Language Assessment Quarterly*, *10*, 196–218. https://doi.org/10.1080/15434303.2012.721423

Xie, Q., & Andrews, S. (2013). Do test design and uses influence test preparation? Testing a model of washback with Structural Equation Modeling. *Language Testing*, *30*(1), 49–70. https://doi.org/10.1177/0265532212442634

Ye, T., Shao, J., Yi, Y., & Zhao, Q. (2023). Toward better practice of covariate adjustment in analyzing randomized clinical trials. *Journal of the American Statistical Association*, *118*(544), 2370–2382.

Zhang, W., & Bray, M. (2020). Comparative research on shadow education: Achievements, challenges, and the agenda ahead. *European Journal of Education*, *55*(3), 322–341. https://doi.org/10.1111/ejed.12413

Zhang, Y. (2013). Does private tutoring improve students' National College Entrance Exam performance?—A case study from Jinan, China. *Economics of Education Review*, *32*, 1–28. https://doi.org/10.1016/j.econedurev.2012.09.008

Zuman, J. P. (1987). *The effectiveness of special preparation for the SAT: An evaluation of a commercial coaching school*. Harvard University.

Authors

ZHANXIN HAO is a postdoctoral research fellow (Shuimu Scholar) at Tsinghua University, China. Email: zhanxin_hao@yeah.net. Her research focuses on education assessment and AI in education, such as student-AI interaction patterns as well as its long-term effects on student learning, motivation, and socioemotional development.

JO-ANNE BAIRD is a professor of educational assessment at the University of Oxford, UK. Email: jo-anne.baird@education.ox.ac.uk. Her research interests are in educational assessment, including system-wide structures and processes, examination standards, marking, and assessment design.

YASMINE EL MASRI is associate director for research at Ofqual and an honorary research fellow at the University of Oxford. Email: yasmine.elmasri@education.ox.ac.uk. Her research interests focus on the role of language in assessment, science education, international large-scale assessments, and the design and delivery of computer-based assessments.

KIT DOUBLE is a senior lecturer and a DECRA fellow at the School of Psychology, University of Sydney, Australia. Email: kit.double@sydney.edu.au. His research focuses on educational psychology, cognition, and individual differences.