Educational Psychology Review (2020) 32:791–814
https://doi.org/10.1007/s10648-020-09520-6

INTERVENTION STUDY

Learning Mathematics Problem Solving through Test Practice: a Randomized Field Experiment on a Global Scale

Francesco Avvisati[1] · Francesca Borgonovi[1,2]

Published online: 4 February 2020

Abstract
We measure the effect of a single test practice on 15-year-old students' ability to solve mathematics problems using large, representative samples of the schooled population in 32 countries. We exploit three unique features of the 2012 administration of the Programme for International Student Assessment (PISA), a large-scale, low-stakes international assessment. During the 2012 PISA administration, participating students were asked to sit two separate tests consisting of problem-solving tasks. Both tests included questions that covered the same internationally recognized and validated framework for mathematics assessment. Students were randomly assigned in the first, 2-h-long test to one of three test versions containing varying amounts of mathematics, reading, and science problems. We found that the amount of mathematics problems in the first test had a small positive effect on mean mathematics performance on the second test, but no effect on general reasoning and problem-solving ability. Subject-specific effects of test practice on subsequent test performance were found over both short lags (same day) and medium lags (1–7 days). The learning gains ascribed to mathematics problem-solving practice were larger for boys than for girls.

Keywords Retrieval practice · Large-scale international assessments · Mathematics · Problem solving · Testing

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10648-020-09520-6) contains supplementary material, which is available to authorized users.

✉ Francesca Borgonovi
f.borgonovi@ucl.ac.uk

[1] Directorate for Education and Skills, Organisation for Economic Co-operation and Development, 2 Rue André Pascal, 75016 Paris, France

[2] Department of Social Science, Institute of Education, University College London, 55-59 Gordon Square, London WC1H 0NU, United Kingdom

School tests have become increasingly prominent in education research and policy in many countries, sparking intense public scrutiny over their intended and unintended consequences (Coburn, Hill, & Spillane, 2016; Marsh, Roediger, Bjork, & Bjork, 2007; McNeil, 2000; Smith, 1991; Rothman, 2011; Tienken & Zhao, 2010; Vinovskis, 2008). High-stakes tests are used, for example, to certify students' acquisition of skills and knowledge and to award access to selective educational and career opportunities (Madaus, 1988). They are also a key component of accountability systems and are used to monitor the performance of teachers and schools. Low-stakes tests, including practice tests, are used by teachers to gather information about learners' progression and to prepare students for high-stakes test. While practice tests can promote some learning (such as learning how to allocate time efficiently on the test, or helping students memorize facts and procedures), the use of tests in today's schools is often criticized by parents, teachers, and educators for promoting only narrow learning that is deemed to be irrelevant in the "real world," narrowing the curriculum (Amrein & Berliner, 2002; Nichols & Berliner, 2007; Watanabe, 2007). Teachers who "teach to the test" and students "who learn for a test" are, in public discourse, considered to divert valuable time resources from learning (Crocco & Costigan, 2007; Nelson, 2013).

Research indicates that the impact of tests on the curriculum and classroom practice depends on the characteristics of the tests. At the level of classrooms, schools, or entire systems, it has been shown that curricular content can narrow as a result of testing, subject area knowledge can become fragmented into test-related pieces, and teachers can increase the use of teacher-centered pedagogies as a result of testing. However, tests have also led to curricular content expansion, the integration of knowledge, and more student-centered, cooperative pedagogies. The effect depends on the characteristics of the test and how tests are integrated in classroom practice (Au, 2007).

At the level of individual learners, findings supported by psychological research suggest that tests can be powerful ways to learn and may serve as useful aids to promote students' ability to apply principles and procedures to new situations. In particular, there is consistent experimental evidence that administering retrieval tests after a study period enhances individuals' ability to retain and recall information (Adesope, Trevisan, & Sundararajan, 2017; Carpenter & Delosh, 2006; Carrier & Pashler, 1992; Kang, McDermott, & Roediger, 2007; Karpicke & Blunt, 2011; Karpicke & Roediger, 2008; Roediger & Karpicke, 2006).

Memorization enables students to store information with the aim of subsequent reproduction (Entwistle & McCune, 2004). The existence of a direct effect of testing indicates that sitting a test that requires the retrieval of previously studied material can be an effective memorization strategy. However, because students are also required to master a range of learning strategies beyond memorization (OECD, 2010; Pritchard, 2013; Rubin, 1981; Weinstein, Ridley, Dahl, & Weber, 1989), the direct effect of testing, while important, is of only partial interest to education practitioners (Rohrer, Taylor & Sholar, 2010; Wooldridge, Bugg, McDaniel, & Liu, 2014).

In addition to recalling definitions and understanding key concepts, students must learn when and how to apply principles and procedures in new settings (Dirkx, Kester, & Kirschner, 2014). Common pedagogical styles that are considered to foster the development of students' ability to apply principles and procedures to new situations are cognitive activation strategies, student-centered instruction and inquiry, and task-based learning (Bietenbeck, 2014). It is therefore important to evaluate if a testing effect can be detected when the criterial test that is used to measure outcomes requires a novel demonstration of learning (Salomon & Perkins, 1989).

Studies have indeed indicated that tests consisting of retrieval tasks can facilitate the encoding of new information and its successive retrieval (Pastötter & Bäuml, 2014;

Szpunar, McDermott, & Roediger, 2008). A rapidly growing area of research examines the extent to which tests consisting of retrieval tasks can have transfer effects and facilitate the application of knowledge in tests consisting of new, different tasks (Barnett & Ceci, 2002). Recent comprehensive reviews (Carpenter, 2012; Pan & Rickard, 2018) have applied the taxonomy of transfer effects proposed by Barnett and Ceci (2002) to identify the extent to which transfer effects occur, if at all, in different contexts and settings (as defined by time, knowledge domains, and test format).

The evidence for a direct effect of testing is robust, with a large number of studies indicating that retrieval practice has a large effect on fact retention (Adesope, Trevisan, & Sundararajan, 2017). Proponents of transfer-appropriate processing (Morris, Bransford & Franks, 1977) suggest that transfer effects will be stronger if the cognitive processes invoked during the practice test are similar to those invoked during the criterial test. The evidence on transfer effects is growing but remains limited, especially when far transfer (i.e., transfer that involves extensive and/or multiple differences in context) rather than near transfer (minor differences in conditions) is considered.

The review by Pan and Rickard (2018) includes evidence from 67 published and unpublished studies on transfer effects of retrieval practice, reporting findings from 122 experiments for an overall study population of just over 10,000 individuals. The review suggests, for example, that in situations involving practice tests, transfer of learning is greatest across test formats (e.g., from a free recall test to a multiple-choice test of the same information) and from memorization to application and inference questions; it is weakest to rearranged stimulus-response items, to untested materials seen only during initial study, and to problems involving worked examples.

The limited number of studies that have been conducted on transfer effects is likely to be the result of the greater complexity of designing such studies when compared to designing studies of the direct effect of testing.

First, because in the case of far transfer, effects are expected to be smaller than for the direct effects of testing, samples have to be considerably larger and more studies have to be conducted for individual studies or for meta-analyses to have adequate power. Small effect sizes can be due to the fact that practicing the retrieval of some information may cause test takers to forget other related information leading to small estimated effects (Anderson, Bjork, & Bjork, 1994; Storm & Levy, 2012) as well as the fact that the link that exists between the material participants are exposed to in the practice test and in the criterial test is less direct in transfer-effects than that in direct-testing-effects studies (with the link being larger in near transfer than in far transfer).

Second, the identification of transfer effects would ideally require that some study participants are administered a "placebo test" in the practice session (i.e., a test that lets student practice test-taking skills, but not exercise the relevant content knowledge), and that the outcome test session include "pseudo-outcomes" (i.e., outcomes on which no effect is expected, given the hypothesized mechanism of transfer). Such a design ensures that transfer and not recall is captured by estimated effects, but also increases sample size requirements.

The present study seeks to contribute to the rich literature on testing effects using data from two field experiments embedded in the 2012 round of the Programme for International Student Assessment (PISA). PISA is an international large-scale assessment that has been administered to samples of 15-year-old students every 3 years since 2000. In each round, a minimum sample of 4500 students per participating country take part in the test and in 2012 over 60 countries participated in PISA. PISA has a global coverage, although few countries from Africa

participated in the study until the most recent round in 2018. The test is administered in a wide variety of cultural, linguistic, and social contexts and stringent technical standards are implemented to ensure comparability (OECD 2014a). Our contribution is threefold.

First, we distinguish between subject-specific learning effects and the improvement of test-taking strategies. In typical direct-testing-effects studies, the treatment consists of a retrieval session while the control is represented by a restudy session. In contrast, in our first experiment (study 1), we compared the mathematics problem-solving ability of students measured on a second test after they all sat a 2-h practice test. The difference between different groups of students was that some were asked to solve up to 1.5 h of content-relevant material (mathematics questions) while others were asked to solve as little as 0.5 h worth of content-relevant material and as much as 1.5 h of content-irrelevant material (comprising a range of science and text comprehension questions). In other words, our control group was administered a placebo test consisting mainly of content-irrelevant test questions. In a second experiment, we further compared similar experimental groups on a second test that did not contain mathematics questions, to ensure that eventual transfer effects of a greater amount of mathematics practice (or exposure) uncovered in study 1 were domain specific.

Second, by using large, representative samples from the schooled population of 15-year-olds in 32 countries worldwide, we achieved adequate power to capture small-to-medium effect sizes. Third, we exploited a field experiment based on authentic educational material to identify the effect of test practice on mathematics problem-solving performance: this has great potential to improve the ecological validity of the study and enhance the ability to draw implications for school and classroom practice.

Theory

Bjork (1994, 1999) defined situations that promote long-term retention and transfer of knowledge at the potential expense of immediate performance as desirable difficulties. Desirable difficulties considered in the literature include distributed practice, varying conditions of practice, contextual interference, and, crucial for our study, testing (also referred to as retrieval practice) (Roediger & Karpicke, 2006; Roediger & Butler, 2011).

The studying of retrieval practice has been largely an empirical effort with the first empirical studies dating back to the early twentieth century (Abbott, 1909) while theoretical work aimed at understanding why a testing effect occurs and why the effect is larger in some conditions lagged somewhat behind (Roediger & Butler, 2011). One of the first prominent attempts to explain why testing effects occur focused on the role of exposure (Thompson et al., 1978). According to this theory, retrieval practice promotes learning because individuals are re-exposed to relevant material. However, subsequent empirical studies which compared test-taking to equivalent amounts of re-study, thereby ensuring similar exposure in both the test and control conditions, continued to find greater effectiveness of retrieval practice over re-study, prompting refinements in the theory (Roediger & Butler, 2011).

Elaborative-retrieval theory and the theory of transfer-appropriate processing define the mechanisms as to why retrieval practice can be considered to be a desirable difficulty and promote learning.

Elaborative-retrieval theory (Carpenter, 2009) maintains that two factors determine the testing effect: spreading activation and semantic elaboration. Spreading activation refers to the process through which retrieval practice strengthens existing retrieval routes and supports

the creation of new retrieval routes. The strengthening of existing retrieval routes and the creation of new ones make it more likely that content will be successfully retrieved in the future (Roediger & Butler, 2011; Pyc & Rawson, 2009). Searching for contents in associative memory networks activates these contents as well as contents associated with it, even if the latter contents are not directly retrieved. Semantic elaboration refers to the amount of retrieval effort directed towards elaboration: greater retrieval effort corresponds to a more extensive reprocessing of the memory trace during retrieval (Roediger & Butler, 2011).

Elaborative retrieval theory explains the existence of a testing effect by considering that answering a series of questions demanding the recall of facts or requiring the application of a given set of principles or procedures will help students consolidate information in long-term memory (Keresztes, Kaiser, Kovács, & Racsmány, 2014) and promote a more integrated mental model that incorporates the target knowledge (Karpicke, 2012). Retrieval practice after initial encoding may also reduce the interference of competing irrelevant memories and reduce the rate of forgetting (Roediger & Karpicke, 2006). Elaborative retrieval theory predicts that the testing effect will be stronger the greater the amount of effort involved during retrieval.

Elaborative retrieval theory considers the effort involved during retrieval practice the factor determining why a testing effect occurs and its strength. By contrast, the theory of transfer-appropriate processing considers the match between the cognitive processes involved during the learning phase (the retrieval test) and those required during the criterial test (Morris et al., 1977). A greater match between the processes involved in the two phases can be expected to be associated with better performance on a final test. According to transfer-appropriate processing, the testing effect occurs because the cognitive processes involved during a practice test are more similar to those required during the final criterial test than those involved in other types of encoding activities, such as, for example, restudy (Roediger & Butler, 2011; Thomas & McDaniel, 2007). The theory of transfer-appropriate processing predicts that the testing effect will be stronger the greater the similarity between the practice and criterial tests in factors such as question format and content evoked (Roediger and Karpicke, 2006).

The theory of disuse (Bjork & Bjork, 1992) provides a comprehensive framework that can be used not only to understand why a testing effect occurs but also to make predictions about which conditions strengthen such effect. The theory distinguishes between storage strength and retrieval strength. Storage strength refers to how permanent a particular memory trace is while retrieval strength refers to how accessible such a trace is. A memory trace is high in storage strength when it is integrated with other representations and is consequently retained over the long term. A memory trace is high in retrieval strength when it is momentarily easily accessed and activated. When retrieval strength is high, short-term performance on a task is enhanced although there may be no appreciable long-term effect on performance. In fact, the theory of disuse maintains that retrieval strength is negatively associated with increments in storage strength: the easier it is to retrieve particular contents (i.e., the less semantic elaboration is involved during retrieval), the less such contents gain in storage strength (because less spreading activation occurs).

According to Bjork's theory of disuse the strength of the testing effect may differ according to the time lag between retrieval practice and subsequent testing events, the spacing of retrieval practice sessions, the mode of retrieval delivery, i.e., whether retrieval practice consists in multiple-choice questions or constructed responses, whether corrective feedback is provided and when such feedback is provided (Adesope et al., 2017).

In line with theoretical predictions, empirical research identifies an advantage of retrieval practice over restudy when the retention interval (the time lag between the treatment condition

and the target test event) is longer than 1 day (Karpicke & Roediger, 2008; Keresztes et al., 2014; Toppino & Cohen, 2009; Wheeler, Ewers, & Buonanno, 2003). The importance of the retention interval as a moderator of the association between practice and performance may be due to the effect of sleep on memory consolidation (Roediger et al., 2010) and the fact that retrieval practice may aid learning by strengthening memory traces and providing additional retrieval routes when individuals search for information in long-term memory (Keresztes et al., 2014).

Research also indicates that spacing retrieval practice over multiple sessions is generally more effective than the administration of a single testing session of the same duration, and that the spacing of sessions is most effective when the lag between sessions is longer and is distributed and spaced through time rather than completed in close succession (Rawson, Vaughn, & Carpenter, 2014; Roediger & Butler, 2011; Lyle et al., 2019).

It has been shown that tests that require participants to provide constructed responses are associated with stronger positive effects than tests that use multiple-choice response formats (McDermott, Agarwal, D'Antonio, Roediger, & McDaniel, 2014). Constructed responses typically require students to engage in more effortful retrieval than multiple-choice questions and effort exerted during retrieval is a factor that importantly explains the variability in the strength of the testing effect (Kang et al., 2007; Pyc & Rawson, 2009). Some even argue that, in the absence of corrective feedback, multiple-choice questions may have a negative effect on learning, particularly among individuals with low levels of baseline knowledge (Butler & Roediger, 2008; McDaniel & Fisher, 1991; Toppino & Brochin, 1989; Toppino & Luipersbeck, 1993). In multiple choice settings, individuals are exposed to a series of possible answers and if they do not know which one is correct, they may preserve, in subsequent testing events, the memory of wrong answers (Fazio, Agarwal, Marsh, & Roediger, 2010).

Finally, although retrieval practice has been shown to be beneficial in the absence of corrective feedback (Adesope et al., 2017), such feedback is associated with an increase in the benefit of testing (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008). The primary mechanism through which testing can promote learning is that tests promote retrieval of information. Therefore, learning effects depend on whether the correct information is retrieved. In the absence of corrective feedback, testing effects can be small or even negative (Kang et al., 2007).

The Present Studies

Elaborative retrieval theory and the theory of transfer-appropriate processing guided our interest in the development of the two studies that are reported in this article. Based on elaborative retrieval theory and the theory of transfer-appropriate processing, we hypothesized that what matters for performance on a criterial test is *the amount of matching effort* expended during practice, i.e., the amount of effort expended on tasks that relate to the same content domain. In study 1, we varied the amount of matching effort expended by varying the amount of content-relevant material in the practice test. In study 2, we varied the amount of matching effort expended by showing the effect of the same practice-test conditions on a criterial test with non-matching content.

More specifically, in study 1, we examined if performance on a criterial test consisting of mathematics tasks was enhanced when participants were involved in practice tests of equal overall length but of varying amount of mathematics content. We hypothesized that students who were exposed to a practice test containing a greater amount of content-relevant material (mathematics tasks) would perform better on a final test compared to students who were

exposed to the same overall amount of testing material but who were exposed to less content-relevant material. In study 2, we examined if performance on a criterial test consisting of non-curricular problem solving tasks was associated with the amount of mathematics content in the practice test. We hypothesized that the amount of mathematics tasks contained in the practice test would not be associated with performance on the final criterial test of problem solving.

The focus on mathematics stems from the fact that mathematics is a cornerstone of curricula in secondary schools irrespective of country or educational track, and that mathematics often acts as a critical filter determining educational and career progression, course choices and occupational paths (Ma & Johnson, 2008).

Study 1

In study 1, we set out to examine if students' skills in solving mathematics problems could be fostered by administering tests consisting of mathematics problem-solving tasks. Similar to existing studies examining the transfer effect of testing, our target task required solving a different set of mathematics problems using principles and procedures that 15-year-olds are expected to master. However, in contrast to existing transfer-effect studies (Butler, 2010; Dirkx et al., 2014), participants in our study practiced distinct problem-solving tasks rather than recall.

We hypothesized that greater practice of mathematics problems was associated with better performance on a standardized mathematics tests. Consistent with the literature on moderators of the relationship between retrieval practice and learning, we expected that the strength of the relationship between the quantity of mathematics practice and individuals' performance on the mathematics test would differ depending on a number of factors.

The first factor considered was the time lag between the retrieval session (practice test) and the criterial test. We hypothesized that practice would be more strongly associated with performance on the criterial test the longer the time lag between the retrieval session and the criterial test. Longer intervals have been considered to increase storage strength (Soderstrom & Bjork 2015) because longer intervals allow not only for the retrieval of related information but may also promote the integration of information with prior knowledge (Lyle et al. 2019).

The second factor considered was test-takers' achievement in mathematics. We hypothesized that the relationship between retrieval practice and performance on the criterial test would be strongest among high-achieving students. We made this hypothesis because participants in our study did not receive corrective feedback after the retrieval practice session; therefore, in the absence of feedback, high-achieving students were more likely to retrieve correct information and to apply the correct principles and procedures during the retrieval session. Furthermore, in our study, we considered how well students performed on a mathematics test designed to assess how well they can apply principles and procedures that they learned in class over the years to solve a range of problems. High-achieving students are students for whom such principles and procedures have high storage strength and, consequently, retrieval practice could more easily stimulate accessibility.

The third factor considered was how anxious students are toward mathematics. We hypothesized that retrieval practice would be more strongly associated with performance among students with low levels of mathematics anxiety. Mathematics anxiety refers the fear of, or apprehension about, mathematics (Ashcraft & Kirk, 2001). The literature documents a strong negative association between mathematics anxiety and mathematics achievement (Foley et al., 2017). Because individuals who experience mathematics anxiety generally avoid

mathematics (Ashcraft & Ridley, 2005), mathematics anxiety is likely to be associated with students' level of exposure to mathematics tasks. Behavioral (Ashcraft & Kirk, 2001; Park, Ramirez, & Beilock, 2014) and fMRI (Lyons & Beilock, 2011, 2012; Young, Wu, & Menon, 2012) studies suggest that mathematics anxiety creates worries that can deplete resources in working memory—a cognitive system responsible for short-term storage and manipulation of information (Miyake & Shah, 1999) that is important for learning and achieving well in math (Beilock & Carr, 2005; Raghubar, Barnes, & Hecht, 2010).

Finally, although there is no established consensus in the extent to which the effectiveness of retrieval practice varies by gender, we examined gender differences since gender gaps are a recurrent focus of the literature on mathematics performance, attitudes toward mathematics, and engagement in mathematics courses and activities (OECD, 2015).

Study 2

In study 2, we examined whether the administration of mathematics problem-solving tasks improved students' domain-general problem-solving skills, such as deductive reasoning, or students test-taking abilities, such as their time management, rather than specific cognitive processes involved in mathematics problem solving (such as formulating problem situations mathematically or using arithmetic procedures).

Theories of transfer-appropriate processing predict that the testing effect depends on the match between the cognitive processes involved during the final criterial test and those activated during practice tests. We therefore hypothesized that greater practice of mathematics problems (as opposed to any other kind of test practice) would be associated with better performance on a test consisting of mathematics problem solving task (study 1) but not with better performance on a test consisting of domain-general problem-solving tasks (study 2).

Methods

In both studies, we relied on comparisons between three groups created by random assignment, with each group taking a different practice test. Practice tests differed in the amount of mathematics material that they contained but all practice tests were characterized by a lack of feedback. Study 1 and study 2 were conducted at the same time, but on different participants. In this section, we describe how study participants were selected and assigned to the three groups, how materials for the tests were developed, the measures included in the data and the methods used for analyzing them and conducting the experimental comparisons.

Several characteristics contribute to the unique experimental setting employed in our study. Our sample is remarkably large, it covers a large number of countries and it is statistically representative of the wider student population in these countries. Our materials are real-world education assessment tasks developed, translated and validated by a large team of internationally recognized experts. A field trial was conducted prior to the main administration (with a different group of students) to ensure the cross-cultural validity and relevance of the test questions. Each of these features greatly enhances the external validity of our experimental method and the conclusions that can be drawn from it.

Participants: Target Population and Exclusions

Our data come from the 2012 edition of the Programme for International Student Assessment (PISA 2012), a large-scale, cross-national assessment of the mathematics, reading, and science performance of 15-year-old students. All cases used in our analyses were extracted from the public-use files for the PISA 2012 computer-based tests, which can be downloaded from http://www.oecd.org/pisa/pisaproducts/pisa2012database-downloadabledata.htm. We included all 32 national samples of countries that took part in the computer-based assessment of mathematics in our study.[1]

PISA participants were selected from the population of 15-year-old students in each country according to a two-stage random sampling procedure, so that weighted samples are representative of students who are enrolled in grade 7 or above and are between 15 years and 3 months and 16 years and 2 months at the time of the assessment administration (generally referred to as 15-year-olds in this article). In the first stage, a stratified sample of schools was drawn (8321 schools in total across the 32 countries; median school sample within countries—207 schools). In the second stage, students were selected at random in each sampled school.

Study 1

In study 1, we focused on students who were assigned to one of the four computer-based forms (out of 24 possible forms) containing only mathematics questions (and thus no questions from other domains). This corresponds to about three to four students per school. Since PISA assigns students to test forms at random, this subset is representative of the wider population of 15-year-old students. In total, 21,013 students included in the PISA sample[2] were assigned to take a 40-min test in mathematics on computers; they define our target population for study 1. We further excluded students from the samples used for our analysis for three reasons:

1. $N = 89$ students (in nine countries) because they used instruments that were adapted for students with special educational needs.
2. $N = 128$ students because they did not attend the first, pen-and-paper test (practice test, T1).
3. $N = 1352$ students because they did not attend the second, computer-based test session (target test, T2).[3]

The final sample size for study 1 is 19,355 students.

[1] We use the word "country" to refer to all national and subnational entities participating in PISA as distinct territories. In most cases, they correspond to entire countries, although in some circumstances they refer to subnational entities or economies. While 65 national samples exist for PISA 2012, only 32 (out of 65) countries that participated in the PISA pen-and-paper assessment opted for conducting a computer-based assessment of students' mathematical skills. Countries may have decided not to participate in the optional computer-based assessment because of both the direct costs (participation in optional assessments determines additional international development costs as well as national implementation costs for countries) and indirect costs (increased administrative burden of the PISA administration) involved. Moreover, some countries did not feel that they possessed the required technical capacity to administer a computer-based assessment in 2012.

[2] Students are included in the PISA sample if they participate in at least one cognitive test session or if they complete a significant fraction of a background questionnaire.

[3] We define attendance in a test as having at least one non-missing response to an assessment task in that test.

Study 2

In study 2, we used a distinct population of students ($N = 19{,}481$), defined by those students from the same 32 countries and 8321 schools who were randomly assigned to forms containing only non-curricular "problem-solving" tasks instead of only mathematics tasks. Similar exclusion rules as for study 1 were followed.

We did not use students who were assigned to the remaining 16 forms (containing mixtures of mathematics, problem solving, and digital reading tasks), nor students who were not assigned to take a computer-based test, in any of our analyses.

Treatment Groups

In order to efficiently cover a wide range of test material in the three subjects, PISA administers, at random, different test forms to different students (OECD, 2014a). We relied on the random allocation of test forms to different students not only to define our target population for the two studies but also to define our treatment and control groups. Table 1 outlines the PISA 2012 assessment design and its organization around 13 paper-based assessment forms.

The 13 paper-based forms differed in the amount of mathematics tasks that they contained. Three groups of forms could be defined in terms of the overall amount of their mathematics content: four forms contained only 0.5 h of testing material in mathematics, three forms contained 1 h, and the remaining six contained 1.5 h of mathematics questions. Correspondingly, in both studies, we defined three treatment arms and allocated students to these based on the test form that they were assigned. Group 1 (G1) comprised the students who had the forms with the smallest amount of mathematics content, group 2 (G2) included students who had the intermediate amount of mathematics, and group 3 (G3) included students assigned to the forms with a majority of mathematics content.

Table 1 How paper-based test forms were used to define treatment groups in study 1 and study 2 that varied in their exposure to math problem-solving tasks

Paper-based form number	Cluster position (30 min)				Group
	1	2	3	4	
Form 2	PS3	PR3	**PM7a**	PR2	1
Form 8	PS2	PR2	**PM4**	PS1	
Form 12	PS1	PR1	**PM2**	PS3	
Form 13	PR1	**PM1**	PS2	PR3	
Form 1	**PM5**	PS3	**PM6a**	PS2	2
Form 3	PR3	**PM6a**	PS1	**PM3**	
Form 9	PR2	**PM3**	**PM5**	PR1	
Form 4	**PM6a**	**PM7a**	PR1	**PM4**	3
Form 5	**PM7a**	PS1	**PM1**	**PM5**	
Form 6	**PM1**	**PM2**	PR2	**PM6a**	
Form 7	**PM2**	PS2	**PM3**	**PM7a**	
Form 10	**PM3**	**PM4**	PS3	**PM1**	
Form 11	**PM4**	**PM5**	PR3	**PM2**	

PM = mathematics, PS = science, PR = reading. P stands for "paper-based". Forms 1 to 7 also exist in an "easy" variant where cluster PM6a and/or PM7a are replaced by clusters PM6b and PM7b, respectively. Bold entries indicates the Mathematics clusters (problem sets)

Forms were assigned to students with equal probabilities. As a result, some 30% of the students were assigned to G1 (their test contained 0.5 h of mathematics and 1.5 h of reading and/or science problems); 24% of students were assigned to G2 (their test contained 1 h of mathematics and 1 h of reading and/or science problems) and the remaining 46% of students were assigned to G3 (their test contained 1.5 h of mathematics and 0.5 h of either reading or science problems). The percentage of individuals in each group varies because there were four test forms in G1, three forms in G2, and six forms in G3.

Procedures

In both studies, all participants completed two tests: a pen-and-paper test consisting of mathematics and non-mathematics tasks, T1; and a computer-based test, T2. In study 1, T2 consisted only of mathematics tasks (T2-m). In study 2, T2 consisted only of non-curricular problem-solving task (T2-ps).

All participants also completed a questionnaire in which they were asked about themselves, their family, and their school experience. Students had up to 2 h to complete the practice test T1 and up to 40 min of time to complete the criterial test T2. The questionnaire was not timed.

The sequence of operations was dictated by PISA international standards: first, students answered the pen-and-paper test (T1), then they completed the 30-min questionnaire, and finally they completed the computer-based test (T2). However, the exact timing of administration was not prescribed and countries and, to some extent, schools were free to choose the timing in order to reduce the costs of test administration (commuting costs for test administrators, etc.) and minimize the disruption to regular school activities.

The assignment of students to different forms in both tests was blind: students did not know which form they would receive before the test session. All paper-based forms had the same cover (except for the form number on the cover page) and the same front-page material, comprising a formula sheet (e.g., the Pythagorean rule) and instructions about answering formats. Until the moment students were allowed to turn pages, at the beginning of the test session and after going through the material in the front page, students were unaware of the amount of mathematics tasks included in the form. Therefore, they could not have selectively studied certain topics and domains before the practice session, based on whether these topics and domains appeared in their practice form or not.[4]

Materials

The pen-and-paper and computer-based tests used test materials that were developed by an international consortium in collaboration with multi-disciplinary expert groups and under the supervision of the PISA governing board, a body of the Organisation for Economic Co-Operation and Development (OECD) composed of representatives of participating countries appointed by their respective governments. Examples of tasks included in the paper- and computer-based tests are available in the Supplementary Material.

[4] General information about the content of PISA tests, such as the test framework and sample tasks from previous PISA assessments, are shared in advance of the test with participating schools and are also available on the web, to ensure informed consent of schools (and depending on national practice and legislation, of students' families) to participate. To our knowledge, students do not prepare specifically for the test. Test frameworks define vast domains of knowledge and typically refer to curricular contents that students have learned over several years of study.

In both studies, the practice test consisted of a mathematics section of varying length (see above, "Treatment groups") and of a reading (text comprehension) and/or a science section; the total length of the practice test was, in all cases, of 2 h. The number of mathematics questions included in T1 varied between 12 questions for test forms in group 1 (see Table 3) and 36 or 37 for test forms in group 3. The total number of questions (including questions in the domains of reading and science) was, in all cases, between 50 (form 11) and 63 (form 8). This number was determined so that the vast majority of students would be able to complete the test within the 2 h allocated (the number varies because some test questions, particularly in reading and science, were longer to solve than others, an aspect that was not necessarily related to their level of difficulty). Indeed, on average across all students, only 1.5% of items were left unanswered at the end of the test (OECD, 2014a, p.236). The content of the mathematics section is described in greater detail below (study 1) and in Supplementary Material.

Study 1

In study 1, the mathematics section of the practice test (paper-based) and the criterial test (computer-based) shared most characteristics; they were developed by the same experts and according to the same framework. The item pool for both tests ensured a balanced representation of all framework aspects (e.g., mathematics content domains, processes, and response formats), with no major imbalances between the two tests. Both tests required interleaved practice of a wide range of mathematical tasks with different kinds of problems presented in an intermixed order (Rohrer, Dedrick & Stershic, 2015). Test developers also paid particular attention to ensuring that, to the extent possible, each of the distinct test forms administered to students did not deviate strongly in these dimensions from the characteristics of the overall item pool (see Supplementary Material for an overview of item types in each test).

No items administered in the practice test were used in the criterial test. The item sets consisted of distinct problems, even though they called on the same type of mathematics knowledge and procedures as tasks in the practice test. For example, simple proportional reasoning—an aspect of the content area "Change and relationships"—was required to solve both task DRIP RATE (question 13), a task included in the practice test, and CD PRODUCTION (question 12), a task in the criterial test (see Supplementary Material). Similarly, in both tests, students were presented with tasks where they needed to read values from a two-dimensional chart (see CHARTS, question 7, and BODY MASS INDEX, question 17). While the scenario and content of the tasks was always different between the tasks in the criterial test and those in the practice test, all major content areas (space and shape; change and relationships; quantity; uncertainty and data) were represented in the criterial test and in each test form used in the practice session (by a minimum of two questions for test forms in group 1, of five questions for test forms in group 2, and seven or more questions for test forms in group 3). In other words, for each task in the criterial test there was always at least one task in the practice test from the same mathematics sub-domain.

In summary, the practice test and the criterial test covered the same content areas and used similar question and response formats. The main difference was the mode of administration (pen-and-paper vs. computer). Both tests consisted entirely of word problems including graphical displays of data and mathematics formulae, with the only format difference between the two being that interactive graphical displays were only included in the criterial test. Furthermore, both tests were targeted at a similar broad range of proficiencies, although lower levels of proficiency were somewhat less well covered by the criterial test.

Study 2

In study 2, the practice test (T1) was identical to study 1, whereas the criterial test (T2-ps) differed from both T1 and from the criterial test used in study 1 (T2-m). The criterial test in study 2 consisted of problem-solving tasks that did not require expert knowledge to solve. The PISA 2012 assessment of problem solving focused on students' general reasoning skills, their ability to regulate problem-solving processes, and their willingness to do so, by confronting students with "complex problems" and "analytical problems" that did not require subject-specific knowledge (e.g., of mathematics) to solve (in "complex problems," students needed to explore the problem situation to uncover additional information required to solve the problem; this interactive knowledge acquisition was not required in "analytical problems") (OECD, 2013; Ramalingam, Philpot & McCrae, 2017). The problem-solving competencies assessed by T2-ps are relevant to all domains (mathematics, reading, and science) assessed in T1, but the ability to solve such problems does not rely on the content and procedural knowledge assessed in any of the three domains present in T1. Example tasks from the assessment of problem solving (T2-ps) can be found in OECD (2014b, pp. 35–44)

Measures

All variables used in the analysis (with one exception highlighted below) are included in the public-use files for PISA 2012, and described in detail in the PISA technical report (OECD, 2014). In particular, we used the following variables in our analyses:

1. PISA test scores, which are included in PISA data sets as multiply imputed measures of proficiency ("plausible values"). PISA test scores are based on item-response-theory scaling procedures and are comparable across students taking different test forms. PISA scores are reported on a standardized scale, with mean 500 and standard deviation 100 across OECD member countries. Therefore, a difference of one PISA point corresponds to an effect size (Cohen's *d*) of 1% of a standard deviation, irrespective of the testing domain.
2. Individual and school-level socio-economic status (the *PISA index of economic, social, and cultural status*), which is based on student responses to the background questionnaire.
3. Age (in months), sex, and grade, which are based on administrative information collected during sampling operations.
4. A binary indicator for students' levels of mathematics anxiety, derived from student answers to the background questionnaire. The anxiety indicator is based on the PISA index of mathematics anxiety, which is standardized to have mean zero across OECD countries. We derive an indicator from this index that distinguishes students with positive values of math anxiety (meaning that their level of anxiety is above average) from the remaining students.[5]

The only information used in this study that is not available in the publicly documented PISA files is the time lag between the pen-and-paper session (T1) and the computer-based session

[5] Because the questions about mathematics anxiety are included in a rotated part of the student questionnaire, the anxiety indicator is available for only about two thirds of the sample. Apart from a small amount of item non-response, missing information about mathematics anxiety is "missing completely at random" due to the randomized assignment of blocks of questions in the student background questionnaire.

(T2). We contacted national centers to obtain session report forms and used these to identify the typical time lag between T1 and T2. Our analysis revealed that most national centers organized the PISA administration so that T1 would take place in the morning and T2 in the afternoon of the same day. However, in Brazil and the Slovak Republic, T2 was administered the day after T1; in Italy, T2 was conducted within a week of T1; and in Macao (China), T2 was administered 3 weeks after T1.

Overview of Analysis

Study 1 and study 2 were both analyzed according to parallel analysis plans.

In order to identify the impact of mathematical-problem-solving practice (the amount of mathematics questions in T1) on subsequent performance in the target task T2-m (study 1) or T2-ps (study 2), we conducted mean comparisons across the three treatment arms within a linear regression framework, to allow for the inclusion of control variables. The treatment consisted of exposing students to a greater dose of problem-solving practice in mathematics at the expense of problem-solving practice in other subjects (problem sets containing reading and science material). Our preferred model measures the treatment as a continuous time variable (the time-equivalent number of mathematics tasks included in the practice, which varies from 0.5 h for G1 to 1 h for G2 and 1.5 h for G2): this assumes that the dose–response relationship is (locally) linear. To identify possible non-linearities, we also compared G2 (1 h) and G3 (1.5 h) to G1 without imposing a functional form on the relationship, by including indicator variables in the regression.

Formally, let (y_i^{T2}) indicate students' performance in T2 and (m_i^{T1}) indicate a "treatment" variable, measuring the intended amount of mathematics tasks in T1. Further, let α_c represent a set of country-specific intercepts and the vector x_i additional baseline controls. We estimated

$$y_i^{T2} = \alpha_c + m_i^{T1'}\beta + x_i'\gamma + \epsilon_i$$

In the above equation, error terms ϵ_i are assumed to be correlated across students attending the same school and are allowed to be correlated within countries as well, but they are assumed to be independent across countries. We took the multi-level, stratified sample design into account by using balanced repeated replication weights, in line with recommended procedures for secondary data analysis in PISA (OECD, 2009). In our preferred specification, the "treatment" variable is a scalar variable measuring the intended amount of mathematics tasks in T1 (m_i^{T1}) in hours; we also ran regressions where m_i^{T1} represents a 2×1 vector with two dummy indicators for group 2 (1 h) and group 3 (1.5 h).

In study 1, we used the same regression framework to compare the demographic and socio-economic characteristics of students in the three groups and to verify the balanced nature of the three experimental groups; in this case, we excluded control variables from the regression.

To identify the moderating effect of baseline student characteristics and of the time lag between tests on our coefficient of interest β, we interacted the treatment variable with indicator variables (or vectors) for the typical time-lag between tests (a country-level variable), for sex, for performance levels in T1, and for levels of self-reported mathematics anxiety.

Results

Study 1

Sample Descriptives and Balancing Tests

Table 2 presents descriptive statistics for study 1 participants. Formal tests confirmed that the small associations of the assignment variable with baseline characteristics, captured by the "beta" coefficient in Table 2, were well within the confidence intervals for random associations. Students' performance on mathematics tasks in the practice session—a proxy for their baseline potential and for the effort exerted in the practice session—was also found to be equivalent across the three groups. In order to maximize our ability to detect small effects of testing, we therefore included controls for performance in T1 in subsequent analyses.

Effect of Math Test Practice on Math Performance

We investigated the effects of mathematics problem-solving practice by looking at how students' results in the target test varied depending on their exposure to mathematics problems in the practice session. Results showed that the greater the proportion of mathematics problems included in the practice session, the better students performed on the target test. In our preferred model, which assumes a linear dose–response relationship, 1 h of additional problem-solving practice in mathematics improved students' ability to solve mathematics problems by about 2% of a standard deviation ($b = 2.29$, $SE = 0.84$, $p = 0.007$) (see model 1 in Table 3). Mean comparisons that do not assume a functional form for the dose–response relationship confirmed that both G2 and G3 outperformed G1 on the target task ($b_{G2\text{-}G1} = 2.40$, $SE_{G2\text{-}G1} = 1.09$; $b_{G3\text{-}G1} = 2.41$, $SE_{G3\text{-}G1} = 0.87$) (model 2 in Table 3). Both differences were significant, and we could not reject the hypothesis of a linear dose–response relationship at conventional levels of significance ($p = 0.178$).

Table 2 Descriptive statistics on baseline variables (study 1 sample)

Variable	Mean	Standard deviation	Beta	Standard error	*p* value
Age	15.779	0.294	0.001	0.006	0.923
Sex (0 = boy, 1 = girl)	0.501	0.500	−0.006	0.010	0.501
ESCS (PISA index of economic, social, and cultural status; set to 0 if missing)	−0.128	0.999	−0.018	0.019	0.349
Average school ESCS (computed on full PISA sample)	−0.133	0.674	−0.008	0.012	0.479
Missing ESCS (0 = not missing, 1 = missing)	0.011	0.106	−0.001	0.002	0.732
Immigrant (0 = non-immigrant, 1 = first- or second-generation immigrant)	0.129	0.335	−0.001	0.006	0.816
Student is above modal grade for 15-year-olds	0.071	0.258	−0.003	0.005	0.452
Student is below modal grade for 15-year-olds	0.237	0.425	−0.014	0.008	0.085
Mathematics performance in T1	499.2	104.4	0.40	2.03	0.843

Beta is the coefficient estimated from a regression of the variable indicated in the first column on the continuous treatment variable

Table 3 The effect of students' exposure to math problems in a practice test on performance in a math test (study 1) and in a problem-solving test (study 2)

Treatment variable:	Study 1 dependent variable: math score in T2-m				Study 2 dependent variable: general reasoning and problem-solving score in T2-ps	
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Math practice (in hours)	2.29**	2.65+			0.95	0.36
	(0.84)	(1.48)			(1.16)	(1.81)
Group 2 indicator			2.40*	1.48		
			(1.09)	(1.90)		
Group 3 indicator			2.41**	2.67+		
			(0.87)	(1.50)		
Control variables:						
Math score in T1 (cubic polynomial)	Included	Not included	Included	Not included	Included	Not included
Demographics	Not included	Included	Not included	Included	Not included	Included
Country dummies	Included	Included	Included	Included	Included	Included
N	19,355	19,355	19,355	19,355	19,481	19,481
p value for beta = 0	0.007	0.074			0.413	0.844
p value for beta2 = 2 × beta1			0.178	0.928		

Demographic control variables include students' age (in months), sex and grade, their socio-economic status, and the average socio-economic status of students in their school. Standard errors in parentheses

Models 1, 2, 5, and 6: the key independent variable is the continuous indicator of exposure to mathematics content in hours. Models 3 and 4: the key independent variables are two dichotomous indicators for group membership (G2 = exposure to 1 h of mathematics content; G3 = G2 = exposure to 1.5 h of mathematics content). Baseline G1 = exposure to 0.5 h of mathematics content

+$p < 0.10$, *$p < 0.05$, **$p < 0.01$

Heterogeneity in Treatment Effects

We investigated the moderating influence of the time lag between the practice test and the target test and of student baseline characteristics on the effects on math performance highlighted in Table 4. All moderation analyses are presented with mathematics performance in T1 as a control variable. Table 4 summarizes the results of moderation analyses.

Time Lag

Effects were positive and significant both for countries in which T1 and T2 were given on the same day ($b_{lag0} = 1.74$, $SE_{lag0} = 0.88$) and for the three countries in which T2 was administered

Table 4 Study 1: how the effect of students' exposure to math problems varies across groups

Dependent variable:	Math score in T2-m			
Sub-group interaction effects:	Model 7	Model 8	Model 9	Model 10
T1/T2 lag (same day; 1–7 days)				
Math practice interacted with lag (same day)	1.74*			
	(0.88)			
Math practice interacted with lag (1–7 days)	7.87*			
	(3.58)			
Sex				
Math practice interacted with sex (girl)		0.34		
		(1.11)		
Math practice interacted with sex (boy)		4.12**		
		(1.41)		
Math anxiety				
Math practice interacted with anxiety level (low)			2.89	
			(1.78)	
Math practice interacted with anxiety level (high)			2.16	
			(1.43)	
Math score in T1 (five performance bands)				
Math practice interacted with (lowest performance: level 1)				−1.86
				(2.28)
Math practice interacted with (level 2)				1.99
				(1.56)
Math practice interacted with (level 3)				2.82*
				(1.23)
Math practice interacted with (level 4)				5.09*
				(2.09)
Math practice interacted with (highest performance: level 5)				4.73
				(3.07)
Control variables:				
Math score in T1 (including squared and cubed)	Included	Included	Included	Included
Sex dummy	Not included	Included	Not included	Not included
Anxiety dummy	Not included	Not included	Included	Not included
Country dummies	Included	Included	Included	Included
N	18,833	19,355	12,665	19,355
p value for equal effects	0.103	0.045	0.766	0.327

In all models, math practice is a continuous variable expressed in hours, taking values between 0.5 (for participants in G1) and 1.5 (for participants in G3). Interaction effects show the math-practice effect among the specific sub-groups identified (lag between T1 and T2-m, sex, mathematics anxiety, prior math achievement). Data for Macao (China) are not included in model 7 because the time lag between test sessions (3 weeks) was significantly longer than in all other countries

Standard errors in parentheses

+$p < 0.10$, *$p < 0.05$, **$p < 0.01$

on a different day, but within a week of T1 ($b_{lag1} = 7.87$, $SE_{lag1} = 3.58$) (model 7 in Table 4). While the latter estimate is substantially larger than the former, the large standard errors around these estimates mean that we could not exclude, at conventional levels of significance, that results were equally strong for the two sets of countries ($p = 0.103$). These results clearly indicate that the observed effects were not limited to the hours immediately following the intervention and may have had a longer-term impact on student learning.

Sex

The observed learning effect of mathematics problem-solving practice was significantly larger for boys ($b_{boy} = 4.12$, $SE_{boy} = 1.41$) than for girls, whose estimated effect is close to zero ($b_{girl} = 0.34$, $SE_{girl} = 1.11$). A formal test rejected equal effects ($b_{boy}–b_{girl} = 3.78$, SE = 1.89, $p = 0.045$) (see model 8 in Table 4).

Anxiety

Results presented in Table 4 (model 9) suggest that students who reported above-average levels of math anxiety may have benefited just as much as students with lower levels of math anxiety from test practice. Point estimates were similar for the two groups (none of which reaches statistical significance alone, due, also, to the smaller sample size for this analysis).

Mathematics Proficiency in the Practice Test

Our results were also relatively inconclusive regarding the moderating effect of mathematics achievement. For the purpose of this moderation analysis, students were assigned to one of five proficiency groups depending on their score in mathematics in T1; levels 2, 3, and 4 are the same as the corresponding levels of mathematics literacy identified and described by the mathematics expert group that guided the development of the assessment (OECD, 2014a, p. 297). Level 1 includes all students whose performance lay below the lower score limit for level 2; level 5 includes all student whose performance lay above the upper score limit for level 4. While the point estimates seemed to vary across the performance distribution and were individually significant only for moderate-to-high levels of prior performance (levels 3 and 4), we could not reject uniform effects across the performance distribution, meaning that the observed variation is compatible with the null hypothesis of no variation (model 10 in Table 4). The pattern of statistical significance may reflect the relatively poor coverage of the lower levels of proficiency in the criterial test, which made it hard for low-achieving students to demonstrate any progress (see the section on Materials above and the Supplementary Material).

Study 2

Effect of Math Test Practice on Domain-General Problem Solving

Participants in study 2 were assessed on a criterial test not including any mathematics questions. We found no difference in performance among the 19,481 students who sat a test aimed at measuring general reasoning and problem-solving ability in T2 ($b = 0.95$, $SE = 1.16$, $p = 0.413$) between groups that differed in the amount of mathematics test practice (see models 5 and 6 in Table 3).

Discussion

Our results shed new light on the role of tests in education. Teachers generally use tests as a way to evaluate the skills students have acquired, to certify students' knowledge, or to adapt their teaching to the pace of progress of heterogeneous student populations. They tend to consider tests as assessment tools rather than learning tools (Wooldridge et al., 2014). Learning is usually regarded as the product of information-encoding activities, such as lectures, seminars, and independent study. Tests, on the other hand, are considered to be, at best, neutral events for learning. In fact, concerns over standardized tests are mounting in many countries amidst fears that such tests reduce the time available for encoding activities, and that teachers and students will focus their efforts on what tests measure at the expense of acquiring other valuable knowledge and skills.

The experience of students participating in PISA 2012, an assessment developed to measure students' ability to apply knowledge to real-life problems and situations, demonstrates that even in the absence of corrective feedback, participation in a one-session low-stakes test is not associated with lower performance and, in fact, with a small positive gain in mathematical problem-solving skills. Importantly, the learning effects that we estimate appear to persist over medium time lags (1–7 days).

Crucially, because our design compared students sitting tests that differed only in the amount of content-relevant and content-irrelevant material, our estimates refer to benefits that come on top of improvements in generic test-taking skills that students may acquire by completing a practice test.

Our results indicate that one additional hour spent solving test questions that require students to practice the application of subject-specific principles and procedures was associated with an improved performance of 2% of a SD in students' ability to retrieve and use such principles and procedures to solve new sets of problems. While the effect may appear to be very small according to standard levels first introduced by Cohen (1988), Funder and Ozer (2019) note that when reliably estimated (a condition that our analysis satisfies), what is typically considered to be a very small effect for the explanation of single events can have potentially consequential effects. Moreover, effect sizes need to be evaluated and classified within a frame of reference. In our context, if test practice effects are additive (a condition not rejected by our data), spending longer or repeated test sessions may be associated with larger effects. In fact, evidence from the literature indicates that the effect of retrieval practice is greater when such practice takes place in multiple sessions (Roediger & Butler, 2011). Assuming a linear dose–response relationship, our results suggest that five additional hours spaced across time may be associated with an improved performance of about 10% of a SD. This effect is comparable to the estimated average difference in mathematics performance between males and females at age 15 (OECD, 2015).

Our findings therefore indicate that by practicing problem solving in mathematics, students learn, first and foremost, how to solve mathematics problems and, in line with our hypothesis, that the testing effect depends on the amount of matching effort expended during retrieval practice.

We found that test-practice effects were positive on average, but also that test practice might have widened existing sex differences in mathematics performance because its benefits were larger for boys. This finding should be considered and evaluated given existing gender gaps in mathematics proficiency, particularly among high-achieving students (Hedges & Nowell, 1995; OECD, 2015). Contrary to our hypothesis, we did not find differences across students with

different levels of mathematics anxiety. This result may be due to the fact that the PISA test is low-stakes, meaning that neither performance on the practice test, nor performance on the criterial test, had any consequence for test takers. The absence of a moderating effect for mathematics anxiety may not be generalizable to situations in which tests are consequential for respondents.

Do the performance differences observed in the target task imply that students' mathematics skills improved? While prior studies, with different settings and on different populations, found testing effects similar to those described in study 1 (see for example Adesope, Trevisan, & Sundararajan, 2017 and Pan & Rickard, 2018 for comprehensive reviews), theories of transfer-appropriate processing cannot exclude that the testing effect observed in those studies occurred because students gained test-taking abilities and transferred those from practice tests to target tests, rather than the domain-specific principles and procedures which educators care most about. We developed two studies to test the hypothesis that the testing effect is driven by the amount of matching effort expended during retrieval practice and exclude that students merely became better at solving tests, rather than at mathematics, unlike most studies of testing effects. In both studies, all participants, irrespective of the specific group they were randomly assigned to, sat a 2-h test during the practice session. In study 1, we induced variation in the amount of matching effort expended during retrieval practice by administering a different amount of mathematics questions to different groups of students. Participants in the "control" group, G1,[6] spent less time than the treatment group answering mathematics problems in the 2-h test and dedicated this time to answering a combination of text comprehension and science problems. G1 therefore received a "placebo" treatment that could equally well have taught students general problem-solving and test-taking skills (the amount of effort was the same but such effort was directed at material that did not match the content present in the criterial test). In study 2, we compared the performance of groups that differed in the amount of mathematics test practice on a criterial test that did not include any mathematics questions, but rather questions directed at assessing general problem-solving abilities. We found no significant difference. The learning effects highlighted in study 1 must therefore be interpreted as the effect of solving mathematics problems on mathematics problem-solving performance, above and beyond any effect of test practice on test performance in general.

The literature on testing effects suggests that providing corrective feedback and offering multiple short test-practice sessions over time can significantly amplify the effect of test taking on subsequent performance (McDermott et al., 2014; Roediger & Butler, 2011). Moreover, feedback by teachers can negate stereotype threat affecting girls' performance in mathematics (Cohen, Garcia, Apfel, & Master, 2006).

Our study suggests that lessons that include targeted test-practice sessions bear the promise of improving students' ability to transfer their knowledge of principles and procedures to new, real-world problems, although effects are not large and appear to accrue, in the absence of feedback, only to boys. Future research should aim to identify and compare how the effect of the type of retrieval practice that we studied is associated with intensity and spacing, overall and among key population subgroups. For example, it is important to identify what is the learning gain that is associated, for example, with 10 h of test practice administered in five sessions of 2 h each or in ten sessions of 1 h each. Future research on retrieval practice should also systematically explore differential effects between boys and girls (and men and women),

[6] We refer to the treatment groups as G1, G2, and G3 in both study 1 and study 2, to underscore that these groups received the same "treatment" (the same practice test) regardless of the study. As explained above, however, study 1 and study 2 consist of different participants.

to establish if the differences highlighted in this study can be generalized to other settings. It would be equally important to establish if changes in conditions, for example, the provision of feedback, might ensure that girls benefited from test practice as much as boys. Finally, our study examined test practice in mathematics and it would be important to consider if findings are generalizable to how students acquire proficiency in mastering principles and procedures that are necessary to solve problems in other domains.

Acknowledgments The authors would like to thank participants in seminars at the Organisation for Economic Co-operation and Development, the Paris School of Economics, Education Testing Service and University College London. They would also like to thank the editor, Fred Paas, several anonymous referees as well as the following individuals for providing input, comments and feedback on previous versions of the manuscript: Sola Adesope, Andrew Elliot, Samuel Greiff, Keith Lyle, John Jerrim, Roberto Ricci, Richard Roberts, Henry Roediger, Matthias von Davier, Allan Wigfield, Kentaro Yamamoto. The manuscript was much improved thanks to them and any errors remain our own. F.B. acknowledges support from the British Academy's Global Professorship programme.

Compliance with Ethical Standards

Conflict of Interest F.A. and F.B. work at the Organisation for Economic Co-operation and Development (OECD), the organization responsible for conducting the PISA study. The opinions expressed and arguments employed in this article are those of the authors and do not necessarily represent the official view of the OECD, its member countries or the British Academy.

References

Abbott, E. E. (1909). On the analysis of the factors of recall in the learning process. *Psychological Monographs, 11*, 159–177.

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.

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22*(7), 861–876. https://doi.org/10.1002/acp.1391.

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10*(18). Retrieved on October 25 at http://epaa.asu.edu/epaa/v10n18/

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(5), 1063–1087. https://doi.org/10.1037/0278-7393.20.5.1063.

Ashcraft, M. H., & Kirk, E. P. (2001). The relationships among working memory, math anxiety, and performance. *Journal of Experimental Psychology: General, 130*(2), 224–237. https://doi.org/10.1037/0096-3445.130.2.224.

Ashcraft, M. H., & Ridley, K. S. (2005). Math anxiety and its cognitive consequences: A tutorial review. In *Handbook of mathematical cognition* (pp. 315–327). New York: Psychology Press.

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher, 36*(5), 258–267.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*(4), 612–637.
Beilock, S. L., & Carr, T. H. (2005). When high-powered people fail: Working memory and "choking under pressure" in math. *Psychological Science, 16*(2), 101–105.
Bietenbeck, J. (2014). Teaching practices and cognitive skills. *Labour Economics, 30*, 143–153.
Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge: MIT Press.
Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). Cambridge: MIT Press.
Bjork, R.A., & Bjork, E.L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Hillsdale NJ: Erlbaum.
Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(5), 1118–1133. https://doi.org/10.1037/a0019902.
Butler, A. C., & Roediger, J. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory and Cognition, 36*(3), 604–616.
Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 35*(6), 1563–1569.
Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21*(5), 279–283.
Carpenter, S. K., & Delosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34*(2), 268–276. https://doi.org/10.3758/BF03193405.
Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20*(6), 633–642. https://doi.org/10.3758/BF03202713.
Coburn, C. E., Hill, H. C., & Spillane, J. P. (2016). Alignment and accountability in policy design and implementation: The common core state standards and implementation research. *Educational Researcher, 45*, 243–251. https://doi.org/10.3102/0013189X16651080.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.
Cohen, G. L., Garcia, J., Apfel, N., & Master, A. (2006). Reducing the racial achievement gap: A social-psychological intervention. *Science, 313*(5791), 1307–1310.
Crocco, M. S., & Costigan, A. T. (2007). The narrowing of curriculum and pedagogy in the age of accountability: Urban educators speak out. *Urban Education, 42*(6), 512–535.
Dirkx, K. J. H., Kester, L., & Kirschner, P. A. (2014). The testing effect for learning principles and procedures from texts. *The Journal of Educational Research, 107*(5), 357–364. https://doi.org/10.1080/00220671.2013.823370.
Entwistle, N., & McCune, V. (2004). The conceptual bases of study strategy inventories. *Educational Psychology Review, 16*(4), 325–345.
Fazio, L. K., Agarwal, P. K., Marsh, E. J., & Roediger, H. L. (2010). Memorial consequences of multiple-choice testing on immediate and delayed tests. *Memory & Cognition, 38*(4), 407–418.
Foley, A. E., Herts, J. B., Borgonovi, F., Guerriero, S., Levine, S. C., & Beilock, S. L. (2017). The math anxiety-performance link: A global phenomenon. *Current Directions in Psychological Science, 26*(1), 52–58.
Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science, 2*(2), 156–168. https://doi.org/10.1177/2515245919847202.
Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science, 269*(5220), 41–45.
Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19*(4–5), 528–558. https://doi.org/10.1080/09541440601056620.
Karpicke, J. D. (2012). Retrieval-based learning active retrieval promotes meaningful learning. *Current Directions in Psychological Science, 21*(3), 157–163. https://doi.org/10.1177/0963721412443552.
Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 331*(6018), 772–775. https://doi.org/10.1126/science.1199327.
Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319*(5865), 966–968. https://doi.org/10.1126/science.1152408.

Keresztes, A., Kaiser, D., Kovács, G., & Racsmány, M. (2014). Testing promotes long-term learning via stabilizing activation patterns in a large network of brain areas. *Cerebral Cortex, 24*(11), 3025–3035. https://doi.org/10.1093/cercor/bht158.

Lyle, K. B., Bego, C. R., Hopkins, R. F., Hieb, J. L., & Ralston, P. A. S. (2019). How the amount and spacing of retrieval practice affect the short- and long-term retention of mathematics knowledge. *Educational Psychology Review*, 1–19. https://doi.org/10.1007/s10648-019-09489-x.

Lyons, I. M., & Beilock, S. L. (2011). Mathematics anxiety: Separating the math from the anxiety. *Cerebral Cortex, 22*(9), 2102–2110.

Lyons, I. M., & Beilock, S. L. (2012). When math hurts: Math anxiety predicts pain network activation in anticipation of doing math. *PLoS One, 7*(10), e48076.

Ma, X., & Johnson, W. (2008). Mathematics as the critical filter: Curricular effects on gendered career choices. In H. M. G. Watt & J. S. Eccles (Eds.), *Gender and occupational outcomes: Longitudinal assessments of individual, social, and cultural influences* (pp. 55–83). Washington, DC, US: American Psychological Association.

Madaus, G. F. (1988). The influence of testing on the curriculum. In L. N. Tanner (Ed.), *Critical issues in curriculum: Eighty-seventh yearbook of the National Society for the Study of Education* (pp. 83–121). Chicago: University of Chicago Press.

Marsh, E. J., Roediger, H. L., Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review, 6*, 194–199.

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology, 16*(2), 192–201. https://doi.org/10.1016/0361-476X(91)90037-L.

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied, 20*(1), 3–21. https://doi.org/10.1037/xap0000004.

McNeil, L. M. (2000). *Contradictions of school reform: Educational costs of standardized testing*. New York: Routledge.

Miyake, A., & Shah, P. (1999). *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge University Press.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*, 519–533.

Nelson, H. (2013). *Testing more, teaching less: What America's obsession with student testing costs in money and lost instructional time*. New York: American Federation of Teachers.

Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage: How high stakes testing corrupts America's schools*. Cambridge, MA: Harvard Education Press.

OECD. (2009). *PISA Data Analysis Manual: SPSS* (2nd ed.). Paris: OECD Publishing.

OECD. (2010). *Mathematics teaching and learning strategies in PISA*. Paris: OECD Publishing.

OECD. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. Paris: OECD Publishing.

OECD. (2014a). *PISA technical report*. Paris: OECD Publishing.

OECD. (2014b). *PISA 2012 results: Creative problem solving (volume V): Students' skills in tackling real-life problems*. Paris: OECD Publishing.

OECD. (2015). *The ABC of gender equality in education: Aptitude, behaviour, confidence*. Paris: OECD Publishing.

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin, 144*(7), 710–756.

Park, D., Ramirez, G., & Beilock, S. L. (2014). The role of expressive writing in math anxiety. *Journal of Experimental Psychology: Applied, 20*(2), 103–111.

Pastötter, B., & Bäuml, K.-H. T. (2014). Retrieval practice enhances new learning: The forward effect of testing. *Frontiers in Psychology, 5*, 286. https://doi.org/10.3389/fpsyg.2014.00286.

Pritchard, A. (2013). *Ways of learning: Learning theories and learning styles in the classroom*. Routledge.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60*(4), 437–447. https://doi.org/10.1016/j.jml.2009.01.004.

Raghubar, K. P., Barnes, M. A., & Hecht, S. A. (2010). Working memory and mathematics: A review of developmental, individual difference, and cognitive approaches. *Learning and Individual Differences, 20*(2), 110–122.

Ramalingam, D., Philpot, R., & Mccrae, B. (2017). The PISA 2012 assessment of problem solving. In B. Csapó & J. Funke (Eds.), *The nature of problem solving: Using research to inspire 21st century learning*. Paris: OECD Publishing.

Rawson, K. A., Vaughn, K. E., & Carpenter, S. K. (2014). Does the benefit of testing depend on lag, and if so, why? Evaluating the elaborative retrieval hypothesis. *Memory & Cognition, 43*(4), 619–633. https://doi.org/10.3758/s13421-014-0477-z.

Roediger, H. L., Agarwal, P. K., Kang, S. H. K., & Marsh, E. J. (2010). Benefits of testing memory: Best practices and boundary conditions. In G. M. Davies & D. B. Wright (Eds.), *New frontiers in applied memory*. Brighton: Psychology Press.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20–27.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x.

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 36*(1), 233–239.

Rohrer, D., Dedrick, R. F., & Stershic, S. (2015). Interleaved practice improves mathematics learning. *Journal of Educational Psychology, 107*, 900–908.

Rothman, R. (2011). *Something in common: The common core standards and the next chapter in American education*. Cambridge, MA: Harvard Education Press.

Rubin, J. (1981). Study of cognitive processes in second language learning. *Applied Linguistics, 2*, 117.

Salomon, G., & Perkins, D. N. (1989). Rocky roads to transfer: Rethinking mechanisms of a neglected phenomenon. *Educational Psychology, 24*, 113–142.

Smith, M. L. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher, 20*(5), 8–11.

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science, 10*(2), 176–199.

Storm, B. C., & Levy, B. J. (2012). A progress report on the inhibitory account of retrieval-induced forgetting. *Memory & Cognition, 40*(6), 827–843. https://doi.org/10.3758/s13421-012-0211-7.

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(6), 1392–1399. https://doi.org/10.1037/a0013082.

Thomas, A. K., & McDaniel, M. A. (2007). The negative cascade of incongruent generative study-test processing in memory and metacomprehension. *Memory & Cognition, 35*(4), 668–678.

Thompson, C. P., Wenger, S. K., & Bartling, C. A. (1978). How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 210–221.

Tienken, C. H., & Zhao, Y. (2010). Common core national curriculum standards: More questions and answers. *AASA Journal of Scholarship and Practice, 6*(3), 3–10.

Toppino, T. C., & Ann Brochin, H. (1989). Learning from tests: The case of true–false examinations. *The Journal of Educational Research, 83*(2), 119–124.

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval. *Experimental Psychology, 56*(4), 252–257. https://doi.org/10.1027/1618-3169.56.4.252.

Toppino, T. C., & Luipersbeck, S. M. (1993). Generality of the negative suggestion effect in objective tests. *The Journal of Educational Research, 86*(6), 357–362.

Vinovskis, M. (2008). *From a nation at risk to no child left behind: National education goals and the creation of federal education policy*. New York, NY: Teachers College Press.

Watanabe, M. (2007). Displaced teacher and state priorities in a high stakes accountability context. *Educational Policy, 21*(2), 311–368.

Weinstein, C. E., Ridley, D. S., Dahl, T., & Weber, E. S. (1989). Helping students develop strategies for effective learning. *Educational Leadership, 46*(4), 17–19.

Wheeler, M., Ewers, M., & Buonanno, J. (2003). Different rates of forgetting following study versus test trials. *Memory, 11*(6), 571–580. https://doi.org/10.1080/09658210244000414.

Wooldridge, C. L., Bugg, J. M., McDaniel, M. A., & Liu, Y. (2014). The testing effect with authentic educational materials: A cautionary note. *Journal of Applied Research in Memory and Cognition, 3*(3), 214–221. https://doi.org/10.1016/j.jarmac.2014.07.001.

Young, C. B., Wu, S. S., & Menon, V. (2012). The neurodevelopmental basis of math anxiety. *Psychological Science, 23*(5), 492–501.