Educational Psychology Review (2024) 36:59
https://doi.org/10.1007/s10648-024-09882-1

ESSAY

Effects of Test Anxiety on Self-Testing and Learning Performance

Shaohang Liu[1,2] · Wenbo Zhao[3] · David R. Shanks[4] · Xiao Hu[2,5] · Liang Luo[1,2] · Chunliang Yang[1,2]

Accepted: 29 March 2024 / Published online: 7 June 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract

Practice testing (i.e., practice retrieval) has been established as an effective learning strategy. Uncovering potential factors influencing self-testing usage is a prerequisite to promote its practical use. The present study reports five experiments exploring whether test anxiety (TA) and test stake (1) affect self-testing usage (Experiments 1–5) and (2) influence learning performance through their negative effects on self-testing usage (Experiments 1, 4, and 5). Experiment 1 analyzed data from 459 high school students collected via a survey and found both that TA negatively predicted students' daily use of self-testing and that self-testing usage mediated the negative association between TA and academic performance. The negative association between TA and self-testing usage was further replicated in a laboratory experiment (Experiment 2). Another quasi-experiment (Experiment 3) showed that students were less likely to test themselves when preparing for a high-stake than a low-stake exam. Experiment 4 replicated this finding and additionally demonstrated that a high-stake test led to poorer learning via its negative influence on self-testing usage. Experiment 5 demonstrated that a high-stake test provoked high state anxiety, which then induced avoidance of self-testing and ultimately impaired learning. Overall, these findings demonstrate a negative effect of TA on self-testing usage, in turn leading to poor learning. Practical implications are discussed.

Keywords Test anxiety · Self-testing · Learning performance · Test-enhanced learning · Classroom

Testing (i.e., retrieval practice) is a more powerful learning strategy by comparison with many others, such as restudying (Roediger & Karpicke, 2006), note-taking (Heitmann et al., 2018), and concept mapping (Karpicke & Blunt, 2011). This

Author Note All data and experimental stimuli have been made publicly available via the Open Science Framework at https://osf.io/3czqr/.

Extended author information available on the last page of the article

phenomenon is known as the *testing effect*, the *retrieval practice effect*, or *test-enhanced learning* (Carpenter, 2009; Rawson & Dunlosky, 2012; Roediger & Karpicke, 2006). Testing can facilitate learning via a range of mechanisms. First, hundreds of studies have confirmed that testing can consolidate long-term retention of studied information, known as the *backward testing effect* (for reviews, see Karpicke, 2017; Rowland, 2014; Yang et al., 2021). Second, testing can boost subsequent learning of new information, known as the *forward testing effect* (Ma et al., 2022; Szpunar et al., 2013; for reviews, see Pastötter & Bäuml, 2014; Yang et al., 2018). Third, by comparison with other strategies (e.g., restudying), testing can yield superior transfer learning. That is, taking practice tests not only facilitates memory for factual knowledge, but also boosts knowledge transfer in the service of solving applied problems (Rohrer et al., 2010; for a review, see Carpenter, 2012). Test-enhanced learning has been demonstrated across different types of learning materials (e.g., foreign-translation pairs, text passages, lecture videos), settings (e.g., laboratory and classroom), and populations (e.g., elementary children, middle/high school adolescents, young college students, and older adults) (for reviews, see Rowland, 2014; Shanks et al., 2023; Yang et al., 2021, 2023).

Given that the benefits of testing are substantial, learners would be well advised to actively utilize testing to enhance their learning in a range of settings both inside and outside the classroom (Dunlosky et al., 2013). However, previous studies have, disappointingly, found that learners do not frequently test themselves and commonly assume that testing is just an assessment tool instead of an effective study strategy (Hartwig & Dunlosky, 2012; Geller et al., 2018; Kornell & Bjork, 2007). The substantial benefits of testing and the underemployment of self-testing jointly present a paradox and highlight an important "know-do gap" between research and practice: Although testing effectively enhances learning, this strategy is not as fully utilized as it deserves to be (Yang et al., 2021). Hence, it is of critical importance to explore what factors constrain self-testing usage. Uncovering such factors is a prerequisite to developing practical interventions to promote self-testing usage and to bridge the know-do gap noted above.

Potential Factors Affecting Self-Testing Usage

Previous investigations using questionnaire surveys have explored self-testing usage by asking individuals to report how frequently they self-administer tests in daily learning (e.g., Bartoszewski & Gurung 2015; Biwer et al., 2020, Weissgerber & Reinhard, 2018). Other studies have conducted controlled experiments in which participants are asked to choose among different strategies for each study item (e.g., restudying it vs. taking a practice test on it) in a real learning task (Toppino et al. 2018, Tullis et al., 2018). Both approaches confirm that self-testing is underappreciated and underutilized by learners (Karpicke et al., 2009; Kornell & Bjork, 2007). These findings stimulated recent studies to explore what factors constrain self-testing usage.

A widely studied factor is erroneous metacognitive beliefs about test-enhanced learning. That is, learners tend to lack metacognitive appreciation of the benefits of

testing (Bjork et al., 2013; Rivers et al., 2022; Roediger & Karpicke, 2006). A recent review conducted by Rivers (2021), which aggregated results across 10 questionnaire studies ($N = 4240$), showed that most (52%) students simply regarded testing as an assessment tool to capture their current level of mastery, with only 26% of them considering testing as an effective learning strategy. These findings are consistent with the *testing-as-monitoring* (TAM) hypothesis (Badali et al., 2022), which proposes that learners generally regard testing as an opportunity for diagnosing learning status (or learning progress) rather than a tool for facilitating learning.

Other supporting evidence for metacognitive unawareness of test-enhanced learning comes from laboratory experiments which found that participants provided higher judgments of learning (JOLs; i.e., judgments about the likelihood of remembering a studied item in a later test) to restudied than to tested materials, even though their memory for tested materials was in fact better than for restudied ones (e.g., Kornell & Son, 2009; Tullis et al., 2013). A possible explanation for this metacognitive illusion is that learners tend to construct JOLs (or judgments about strategy effectiveness) based on processing fluency (i.e., "easily learned, easily remembered"; Bjork et al., 2013; Kirk-Johnson et al., 2019; Yan et al., 2016; Yang et al., 2021), and testing (i.e., retrieving information from memory) is typically more mentally taxing and associated with lower levels of processing fluency than other study strategies, such as restudying. Put differently, learners judge testing as an "effortful and blunt" strategy rather than an "effortful and smart" one.

Aside from erroneous metacognitive beliefs, learning context and material type also affect self-testing choices. For example, studies have shown that learners are more likely to test themselves when learning vocabulary than when learning other types of materials (Vaughn & Kornell, 2019), when holding flashcards in their hand than when using a board (within their reach) to recall the materials (Bottiroli et al., 2010), and when studying in a group compared to studying alone (McCabe & Lummis, 2018; Wissman & Rawson, 2016).

Although several factors influencing self-testing have been identified, to our knowledge no research has been conducted to explore whether test anxiety (TA) affects self-testing usage (see below for a detailed discussion about why TA may affect self-testing). Specifically, it is unknown whether students would avoid testing themselves due to extensive anxiety or other uncomfortable feelings aroused by tests.

McDaniel and Einstein (2020) recently developed a framework for study strategy interventions and proposed that learners' study strategy adoption is affected by four critical elements: knowledge, belief, commitment, and plan (KBCP). In terms of knowledge, students should have metacognitive knowledge of what a given strategy is and how it should be implemented. Take self-testing as an example: Students should know what self-testing is (e.g., reciting information from memory is a kind of practice testing but drawing a brain map is not) and how to perform self-testing. In terms of beliefs, students should believe that testing can truly benefit their learning. Regarding commitment, students should have the willingness to implement a given strategy. For example, students may not choose to test themselves if they think testing is too mentally taxing. Students may also lack motivation to test themselves if they think testing is too threatening and anxiety-inducing. In terms of planning,

students should have a clear plan to determine the parameters of a given strategy, such as the timing, scheduling, and frequency of implementing self-testing. Previous studies have explored several factors affecting self-testing from the perspectives of beliefs (e.g., metacognitive awareness of test-enhanced learning) and knowledge (e.g., using flashcards to implement self-testing). However, few studies have examined the role of commitment in self-testing usage. The present study aims to fill this gap by exploring whether TA, as a commitment component, influences self-testing usage.

A conceptual link between TA and self-testing usage can be inferred based on previous findings showing that learners are reluctant to test themselves when testing will potentially yield negative feedback (e.g., poor test performance, retrieval failures). As summarized by Rivers (2021), students are more willing to test themselves (1) when they think the test is easy to perform, (2) when the interval between study and test is relatively short, and (3) when testing is deployed at the end of the entire learning session. A common property of these conditions is that learners are more likely to test themselves when they believe that they can achieve good performance on the test. This means that avoiding negative feedback is likely to be an important factor leading to underemployment of self-testing. Indeed, worry about poor test performance has commonly been considered a core dimension of TA (Liebert & Morris, 1967; von der Embse et al., 2018). Thus, it is reasonable to infer that individuals struggling with high TA may test themselves less frequently during self-regulated learning and overwhelmingly prefer to choose other "non-risky" strategies, such as restudying.

TA and Self-Testing Usage

As a particular type of academic anxiety (for a review, see von der Embse et al., 2018), TA reflects excessive worries, intrusive thoughts, tension, and physiological arousal experienced before, during, and after an assessment (Spielberger & Vagg, 1995). Besides emotional (e.g., feelings of panic) and physiological (e.g., shortness of breath) responses, learners struggling with high TA also frequently engage in a series of cognitive reactions, such as comparing their performance to that of peers and possessing low levels of confidence, which are collectively known as cognitive TA (Cassady, 2004; Cassady & Johnson, 2002). Survey studies have shown that TA is pervasive among students, ranging from 15 to 60% (Putwain & Daly, 2014; Segool et al., 2013; Thomas et al., 2018). High TA exerts detrimental effects on self-efficacy (Adesola & Li, 2018; Putwain & Symes, 2011; Roick & Ringeisen, 2017), academic performance (Hunsley, 1985; Karatas et al., 2013), and subjective well-being (Putwain et al., 2021; Steinmayr et al., 2016).

Previously, researchers viewed TA as a trait construct reflecting one's general tendency to be anxious about tests, which is normally measured by standardized questionnaires or scales, such as the Test Anxiety Inventory (TAI) (e.g., Sarason & Stoops, 1978; Spielberger, 1980; Szafranski et al., 2012). However, recent models challenge the view that TA is only determined by a personality

trait. Instead, a sophisticated cognitive process is engaged when people are confronting stressful events (e.g., high-stake tests). For example, the *transaction model of stress* claims that it is one's evaluation and appraisal of an event that determines the level of experienced stress (Spielberger & Vagg, 1995). Specifically, a stressor triggers two evaluative steps: (a) whether the event has positive or negative influences and (b) whether the individual has the ability to change the stressful situation. This means that TA can also be substantially provoked by situational factors (e.g., high-stake tests), regardless of psychological traits. Indeed, it has been well established that students typically experience higher levels of TA when facing a high-stake than a low-stake exam (Endler & Kocovski, 2001). Besides test stake, some social factors can also affect TA. For instance, according to *social derogation theory*, worries about negative judgments from others (i.e., social worries) are related to high psychological stress and anxiety (Branscombe & Wann, 1994; Symes & Putwain, 2020). Consistent with this theory, it has been shown that informing students that their test scores would be released to peers or classmates, compared to being kept confidential, makes them report higher levels of stress and anxiety (Wenzel & Reinhard, 2021).

In educational settings, students often need to study a mixed set of important (high-stake) and less important (low-stake) contents when preparing for a later course exam. For instance, in a given class, some contents are more important and will be tested in the final course exam (i.e., high-stake contents), whereas others are less important and will not be tested (i.e., low-stake contents). It is possible that students select different strategies to study low- and high-stake information due to differential evaluation and appraisal of the stake level. According to the transaction model of stress, students may cognitively worry about poor exam performance involving high-stake content on the final course exam. Additionally, according to the TAM hypothesis, they typically regard testing as a monitoring (rather than learning) tool. Hence, cognitive TA may reduce learners' willingness to take self-tests on high-stake content.

Previous research typically explored the influences of TA on learning through two approaches (e.g., Hinze & Rapp, 2014; Wenzel & Reinhard, 2021). First, research has probed the relationship between individual differences in TA and students' study behaviors. Accordingly, the present study asks whether individual differences in TA predict individual differences in self-testing usage. Secondly, previous studies also frequently explored whether test stake affects learners' study behaviors. Based on this, the present study asks whether students are less likely to test themselves when preparing for a high-stake exam than for a low-stake exam. Past research has already established that cognitive TA (e.g., cognitive evaluation and appraisal of an upcoming test) can affect learners' beliefs and behaviors outside the test context (e.g., during the test preparation phase; Cassady, 2004; Cipra & Müller-Hilke, 2019; see Yang et al., 2023, for a review). Accordingly, we predict that students may prefer to use other less risky study strategies (e.g., restudying) to prepare for high-stake exams.

TA and Academic Performance

Numerous studies have demonstrated that TA is negatively related to academic performance (for reviews, see von der Embse et al., 2018; Robson et al., 2023). Theoretical explanations of this negative association can be divided into two clusters: (1) interference-based theories and (2) non-interference-based theories. The core assumption of interference-based theories is that TA interferes with cognitive processing during the test phase, in turn leading to poor academic assessment performance. For example, the *processing efficiency theory* posits that although in some situations TA can stimulate motivation and have a positive impact on performance, individuals must devote a large amount of their cognitive resources to buffer the uncomfortable feelings associated with high TA, which in turn leads to a reduction in the storage and processing capacity of working memory and ultimately impairs task performance (Eysenck & Calvo, 1992). Similarly, the *attentional control theory* claims that task performance is dependent on the level of balance between goal-directed and stimulus-directed attention (Eysenck et al., 2007). High TA can disrupt this balance by forcing individuals to pay more attention to threatening stimuli, leaving less attention to accomplish task goals. This imbalance subsequently interferes with the central executive system (i.e., working memory) and impairs performance.

Different from interference-based theories which propose a causal effect of TA on poor test performance during the test phase, non-interference-based theories assume that TA is a consequence of deficits in knowledge and skills (Culler & Holahan, 1980; Kirkland & Hollandsworth, 1980). Notably, Theobald et al. (2022) recently provided evidence that TA does not predict final exam performance when students' knowledge levels were controlled as a confounding variable. Specifically, even though performance statistically dropped from low-stake mock exams to a high-stake final exam, this reduction was not predicted by TA when students' knowledge levels (represented as mock exam performance) were controlled. Additionally, Theobald et al. (2022) found that TA negatively predicted mock exam performance. These findings together suggest that TA does not affect how well individuals perform a test but instead affects how much they learn when preparing for the test. Relatedly, previous studies on the "learning-testing cycle" showed that students with high TA reported poorer study skills, rated tests as more threatening, and prepared less effective test notes, suggesting that TA hinders learning efficiency during the test preparation phase (Cassady, 2004; Cipra & Müller-Hilke, 2019). These findings also suggest that TA (especially, cognitive TA) can influence beliefs and study behaviors outside the test context.

Although research has provided comprehensive explanations for the negative association between TA and academic performance, to date, no study has explored whether TA impairs academic performance through affecting self-testing usage. Nonetheless, previous investigations have shed some light on this issue by showing that individuals with high TA cognitively view testing as more threatening

(Cassady, 2004). Additionally, it has been found that self-testing usage positively predicted students' academic performance (e.g., Hartwig & Dunlosky, 2012; Stewart et al., 2014). Furthermore, Putwain et al. (2012) reported that perceived control can mitigate the negative effect of TA on examination performance. Putwain and colleagues concluded that students with high perceived control and low TA are more able to deploy their study competencies (e.g., using more effective study strategies) to achieve superior learning performance. These findings jointly establish a theoretical inference that high TA may impair academic performance through reducing self-testing usage. Consistent with this hypothesis, the attentional control theory proposes that whether TA impairs task performance (e.g., academic performance) is at least partially dependent on whether it affects the usage of effective strategies (e.g., self-testing) to compensate for the negative influences of TA (Eysenck et al., 2007). Hence, avoidance of self-testing, induced by high TA, may be one of the mechanisms underlying the negative effect of TA on performance.

Overall, based on previous findings (e.g., Cassady, 2004; Hartwig & Dunlosky, 2012) and relevant theoretical accounts, we predict that TA may impair long-term learning through reducing self-testing usage. A further aim of the present study is to empirically test this hypothesis.

TA and Test-Enhanced Learning

Although self-testing, as an effective learning strategy, has been advocated in a range of empirical studies and systematic reviews (e.g., Yang et al., 2021), an important question is whether TA moderates the magnitude of test-enhanced learning (i.e., whether individuals with high TA benefit less or more from practice testing than those with low TA). The answer appears to be negative: Many studies have reported that the benefits of test-enhanced learning are resistant to TA. For instance, Yang et al. (2021) recently conducted a meta-analysis to quantify the magnitude of the testing effect in the classroom. In this meta-analysis, 295 effects were divided into two categories (high-stake vs. low-stake quizzes) according to whether or not class quiz performance contributed to students' final course grades. The results showed that both high- and low-stake quizzes reliably boosted students' academic performance, and more importantly, stake level did not moderate the magnitude of test-enhanced learning. Consistent findings were observed by Yang et al. (2020), who observed in a sample of over 1000 participants that neither trait nor state TA predicted the magnitude of test-enhanced learning (for related findings, see Myers et al., 2021; but also see Tse & Pu, 2012). Although Tse and Pu (2012) reported that students with high TA and low working memory capacity benefit less from retrieval practice, this finding was not replicated by their subsequent study which showed no relation between TA and the magnitude of test-enhanced learning (Tse et al., 2019).

As mentioned above, it seems that TA does not impact the extent to which practice testing benefits learning. However, here, we argue that TA may be a crucial factor affecting self-testing usage, which then influences learning outcomes. In applied learning settings, tests are normally implemented in two ways. One approach employs tests as an instructional strategy, and students are *forced* to take tests (such as class quizzes), as is done by the majority of previous studies in which participants had to undertake practice tests in the test condition. However, aside from taking tests passively as a task requirement or obligatory class activity, another more common case is that students choose to test themselves in a *self-regulated learning* mode, as most learning activities occur without formal instructions and supervision. Thus, although TA does not directly modulate the magnitude of test-enhanced learning (Myers et al., 2021; Tse et al., 2019; Yang et al., 2020), individuals struggling with high TA may have to pay an "opportunity cost" due to avoidance of self-testing. Hence, TA may indirectly affect learning via its detrimental effect on self-testing usage. In consideration of this, the present study also aims to explore if TA exerts an indirect effect on learning performance by affecting self-testing usage during self-regulated learning.

Overview of the Present Study

As discussed above, we identified two important issues underexplored (or unexplored) in previous studies. First, the relationship between TA and self-testing usage has not yet been probed. The present study explored this question through investigating (a) whether students with high TA are more reluctant to test themselves than those with low TA and (b) whether students engage in self-testing less frequently when preparing for a high-stake test than when preparing for a low-stake one. The documented findings may provide evidence supporting the role of commitment in study strategy use, as proposed by the KBCP framework (McDaniel & Einstein, 2020). Second, we aim to explore whether TA (or test stake) impairs learning through its negative influences on self-testing usage. Evidence on this issue may clarify the mechanisms through which TA undermines learning performance (Theobald et al., 2022).

To achieve these research aims, we conducted five experiments and utilized a range of research approaches. Specifically, Experiment 1 analyzed data from a survey study to ask if there is a negative relationship between TA and self-testing usage and if TA impairs academic performance via its negative influence on self-testing usage in daily learning settings. Experiment 2 probed the relation between TA and self-testing usage in a self-regulated learning task. Experiment 3 was a classroom quasi-experiment conducted to examine whether test stake affects self-testing choices when students are preparing for a course exam. Experiment 4 further investigated whether avoidance of self-testing induced by high-stake tests impairs learning. Experiment 5 conceptually replicated Experiment 4's main findings in a between-subjects design, in which we directly manipulated test stake between students in different classes.

Experiment 1

Experiment 1 aimed to explore whether individual differences in TA predict individual differences in self-testing usage and whether TA affects academic performance through its negative influence on self-testing usage.

Method

Participants

Data from 471 high school students (i.e., grade 11) were collected in a school in Northwestern China. In the final data analyses, we excluded participants who (a) did not finish the questionnaires or academic assessment (10 students) or (b) responded with the same answers (i.e., constant responses) to all questionnaire items (2 students). After exclusion, data from 459 students (239 female; $M_{age} = 17.12$, $SD = 0.23$) were available for the final analyses. All participants gave informed consent to participate.

All experiments reported in the present study were approved by the Ethics Committee of the Faculty of Psychology, Beijing Normal University.

Materials and Procedure

Participants completed multiple questionnaires in the classroom under supervision by a pre-trained teacher. Trait TA was measured by the Chinese version of the TAI (Spielberger, 1980; Yue, 1996). The scale consists of 20 items (e.g., "*During tests I feel very tense*") and participants responded to each item on a 4-point Likert scale: (1) *almost never*, (2) *sometimes*, (3) *often*, and (4) *almost always*. Higher average scores indicate a higher level of TA. Cronbach's α for the present sample was 0.81, indicating acceptable internal consistency. The average reported TA was $M = 2.30$ ($SD = 0.50$).

Self-testing usage was measured by the item "*When I am studying, I ____ test myself*." Participants reported their self-testing frequency on a scale from 1 (*never*) to 7 (*almost always*). The average reported self-testing frequency was $M = 3.96$ ($SD = 1.88$).

Academic performance was assessed by a mid-term examination undertaken 1 week after students completed the questionnaires. Academic performance in six subjects, including Chinese, Math, English, Physics, Chemistry, and Biology, was assessed in the mid-term examination, with assessment scores ranging from 0 to 750 points. Specifically, each of the Chinese, Math, and English tests had a maximum score of 150 points. Each of the Physics, Chemistry, and Biology tests had a maximum of 100 points. The schema of this examination was consistent with the Chinese National Higher Education Entrance Examination. The average exam score was $M = 434.31$ ($SD = 42.49$).

Results

Data analyses were performed via JASP 0.16.4 (JASP Team, 2023), with all parameters set as default. For all experiments in the present study, we provide both frequentist and Bayesian statistics.

As expected, there was a negative correlation between TA and self-testing usage, $r = -0.19$, $p < 0.001$, $BF_{10} = 648.88$, suggesting that students with high TA indeed test themselves less frequently during daily learning. Replicating the classical finding, there was a negative correlation between TA and academic performance, $r = -0.38$, $p < 0.001$, $BF_{10} > 1000$. Additionally, consistent with previous findings (e.g., McAndrew et al., 2016; Sotola & Crede, 2021), self-testing usage positively correlated with academic performance, $r = 0.17$, $p < 0.001$, $BF_{10} = 106.89$, suggesting that the more frequently students test themselves, the superior academic attainment they achieve.

To explore whether self-testing usage mediates the negative association between TA and academic performance, a mediation analysis was conducted, with TA as the independent variable, self-testing usage as the mediator, and academic performance as the dependent variable. The total effect of TA on academic performance was significant, $c = -31.73$, 95% CI $[-38.89, -24.58]$, $p < 0.001$. Critically, and as shown in Fig. 1, TA negatively predicted self-testing usage, $a = -0.73$ $[-1.06, -0.39]$, $p < 0.001$, and self-testing usage positively predicted academic performance, $b = 2.36$ [0.42, 4.30], $p = 0.02$. Most importantly, the indirect effect through self-testing usage was significant, $a*b = -1.71$ $[-3.33, -0.10]$, $p = 0.04$, consistent with our a priori hypothesis that TA affects learning performance partially through its negative influence on self-testing usage. The direct effect was also significant, $c' = -30.02$ $[-37.23, -22.77]$, $p < 0.001$, suggesting that the negative effect of TA on academic performance survives after controlling for its indirect effect through self-testing usage.

Discussion

Experiment 1 confirmed the prediction that students with high TA test themselves less frequently in daily learning. More importantly, the mediation results showed

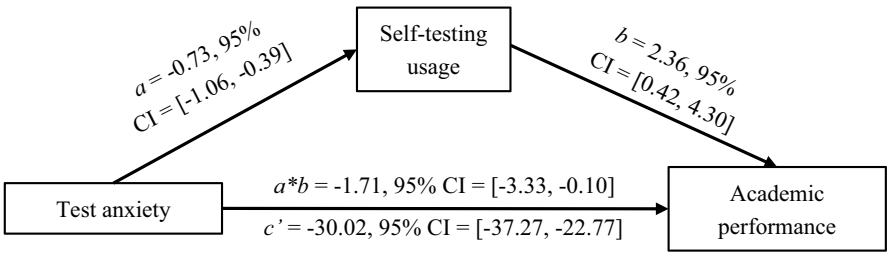

Fig. 1 Mediation model on academic performance in Experiment 1

that TA impairs academic performance partially via its negative influence on self-testing usage.

Two limitations of Experiment 1 should be acknowledged. First, the correlational findings cannot be used to infer a causal relationship between TA and self-testing usage or to directly determine whether the negative influence of TA on self-testing usage further leads to academic deficits. Secondly, Experiment 1 only measured self-testing usage via a single survey item, which means that the observed findings might suffer from measurement error. Considering these limitations, the following experiments were conducted to further explore the effect of TA on self-testing usage and its indirect effect on learning via self-testing usage.

Experiment 2

Experiment 2 was a laboratory experiment conducted to conceptually replicate the negative relation between TA and self-testing usage observed in Experiment 1. Different from Experiment 1 in which self-testing usage was self-reported and measured by a single survey item, Experiment 2 measured self-testing usage in a self-regulated learning task. Specifically, participants were asked to study 30 image-name pairs of human anatomical structures. After studying each image-name pair, they made a strategy choice regarding whether they would like to restudy or take a practice test on it in the next review phase. After they studied all pairs and made strategy choices, they completed a cued recall test on all pairs. The proportions of image-name pairs selected to be tested were taken as the key measure of self-testing usage. According to Experiment 1, we expected to observe a negative relation between TA and self-testing usage.

Note that, although in Experiment 2 participants selected their preferred strategy (restudying vs. testing) for reviewing each item, we did not actually implement their strategy choices afterward (that is, the review phase was actually omitted), because Experiment 2's main research aim was to explore whether TA affects self-testing choices, instead of investigating to what extent self-testing benefits learning. Another reason why we omitted the review phase is that we pre-planned to take participants' final test performance as a control variable when measuring the relation between TA and self-testing choices. Previous studies showed that perceived learning status is a strong predictor of self-testing choices: Learners are willing to test themselves when they believe that they are able to successfully answer test questions (Rivers, 2021). Accordingly, we assume that learners with superior memory ability are more likely to choose self-testing because they would experience fewer recall failures in the tests. Hence, in Experiment 2, we pre-planned to take final test performance (i.e., an indicator of memory ability) as a control variable to better clarify the relation between TA and self-testing usage.

Besides memory ability, we also pre-planned to control the potential confounding effects of effort belief (Blackwell, 2002) and theory of intelligence (Dweck & Leggett, 1988). People have different beliefs about whether (and to what extent) effort improves performance (Blackwell, 2002). Previous studies found that testing, as a kind of "desirable difficulty," is more mentally taxing and requires greater

levels of mental effort than restudying (Karpicke & Roediger, 2007; Kirk-Johnson et al., 2019). Greater levels of mental effort associated with testing may decrease learners' willingness to test themselves, especially for those believing that effort does not improve performance (Koriat, 2008). By contrast, for those believing that effort improves performance, they may be more likely to adopt self-testing. Theory of intelligence (i.e., beliefs about whether one's intelligence is fixed or changeable; see Dweck & Leggett, 1988) may also affect self-testing usage. Individuals who believe that intelligence is incremental tend to adopt more effective and effortful study strategies, whereas those believing that intelligence is fixed are more likely to adopt shallow processing strategies (Costa & Faria, 2018). Hence, the potential confounding effect of intelligence mindset was also controlled in Experiment 2.

Method

Participants

A pilot study ($N = 35$) was conducted to determine the required sample size. The procedure of the pilot study was the same as that in the formal experiment (see below for details). The pilot results showed a negative correlation ($r = -0.56$) between trait TA and self-testing choices. To detect an equivalent effect size with 0.80 power, the required sample size is 35. Considering potential participant attrition (as there were two task sessions separated by a 24-h interval; see below for details), we doubled the sample size.

Finally, 73 undergraduate students were recruited from the participant pool at Beijing Normal University. They confirmed that they were not Biology, Psychology, or Neuroscience majors. Three participants did not finish the day 2 task and were therefore excluded, leaving final data from 70 participants (44 female; $M_{age} = 21.1$, $SD = 2.2$). All participants gave informed consent to participate, reported normal or corrected-to-normal vision, were individually tested in a sound-proofed cubicle, and received monetary compensation.

Materials

Study Stimuli

One hundred images of human anatomical structures were downloaded from a website for medical science education (www.anatomylearning.com). The length of structure names ranged from two to five Chinese characters. To develop appropriate stimuli, another pilot study ($N = 35$) assessed the learning difficulty of each image-name pair. According to the average recall rate of each structure-name pair in the pilot study, the level of difficulty for each item was calculated. Finally, 30 anatomical structure images and their corresponding names were selected to be used in the self-regulated learning task (see Fig. 2). The average recall rate for these structure-name

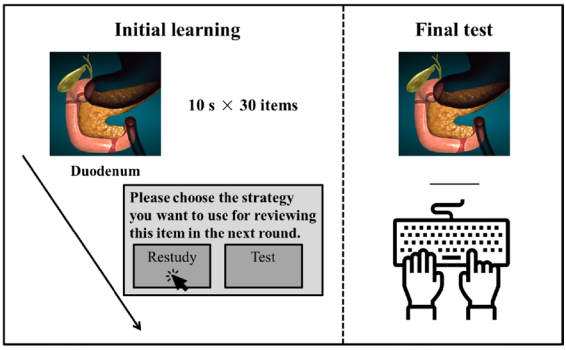

Fig. 2 A trial schema of Experiment 2. Note: participants learned 30 anatomical structures during the initial study phase. After studying each image-name pair, they decided whether to restudy or take a test on that pair during the review phase. In the final test, structure images were presented one by one in random order, and participants were required to recall the corresponding structure names

pairs was 50% ($SD = 9.1\%$). The self-regulated learning task was programmed via *PsychoPy* 2022.2.4 (Peirce, 2007).

TAI

Trait TA was measured by the Chinese version of the TAI. Cronbach's α for the current sample was 0.76. The average reported TA was $M = 2.30$ ($SD = 0.45$).

Effort Belief

Effort belief was measured by a Chinese version of the Effort Belief Scale constructed by Blackwell (2002), composed of four positive items (e.g., "*When something is hard, it just makes me want to work more on it, not less*") and five negative items (e.g., "*If I am not good at a subject, working hard won't make me good at it*"). Participants responded to each item on a 7-point Likert scale (1 = *strongly disagree*, 7 = *strongly agree*). Responses to negative items were reversed. Cronbach's α for the current sample was 0.81. The average score was $M = 3.69$ ($SD = 0.85$), with higher scores representing a stronger belief that effort improves performance.

Theory of Intelligence

Theory of intelligence was measured by the Chinese version of the Theory of Intelligence Scale (Park et al., 2016). The scale consists of eight (four negative and four positive) items (e.g., "*People have a certain amount of intelligence, and they can't really do much to change it*"), and responses were provided on a Likert scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). Cronbach's α for the current sample was 0.74. Responses to negative items were reversed. The average score was $M = 2.86$ ($SD = 0.67$), with higher scores representing a stronger belief that intelligence is incremental.

Procedure

To avoid potential influences of the TA measurement on self-testing choices in the self-regulated learning task, we asked participants to complete the questionnaires and self-regulated learning task on two separate days. On day 1, all participants responded to the questionnaires, including effort belief, theory of intelligence, and TAI.

Twenty-four hours later, participants returned to the laboratory to complete the self-regulated learning task. They were informed that they would study 30 human anatomical structures to prepare for a final memory test. They were also informed that, after studying each image-name pair during the study phase, they would need to make a choice regarding whether they would like to restudy or take a test on it in the next review phase. If restudying was chosen, they would be able to restudy the item for 10 s during the review phase. If testing was selected, during the review phase, the structure image would be shown on screen, and they would have 7 s to recall the name of the structure, following which the correct answer would be shown for 3 s. After the instructions, participants practiced three items to ensure that they understood the experimental procedure.

During the initial study phase, participants learned the 30 image-name pairs one by one in random order. Each image-name pair was presented for 10 s. After studying each pair, participants chose whether they would like to restudy or take a test on that pair during the review phase. They made their choices by selecting one of two options: restudy vs. test. There was no time pressure for them to make strategy choices.

After the initial study phase, participants completed a 2-min distractor task, in which they were asked to solve as many mathematical problems (e.g., *14 + 32 =* __) as they could. After the distractor task, all participants took a final cued recall test. The 30 anatomical structure images were shown one by one in random order, with a blank box shown below each image. Participants were asked to recall the corresponding name for each structure and type their answer into the blank box. There was no time pressure and no feedback in the cued recall test.

Table 1 Correlation matrix in Experiment 2

		TA	EB	ToI	STC
EB	*r*	−.08			
	p (BF_{10})	.49 (0.19)			
ToI	*r*	.29*	.33**		
	p (BF_{10})	.01 (2.81)	.005 (7.12)		
STC	*r*	−.38**	.19	.20	
	p (BF_{10})	.001 (24.63)	.11 (0.51)	.09 (0.59)	
FTP	*r*	−.25*	.03	−.020	.31**
	p (BF_{10})	0.04 (1.23)	.84 (0.15)	.87 (0.15)	.009 (4.15)

TA = test anxiety; *EB* = effort belief; *ToI* = theory of intelligence; *STC* = self-testing choices; *FTP* = final test performance

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

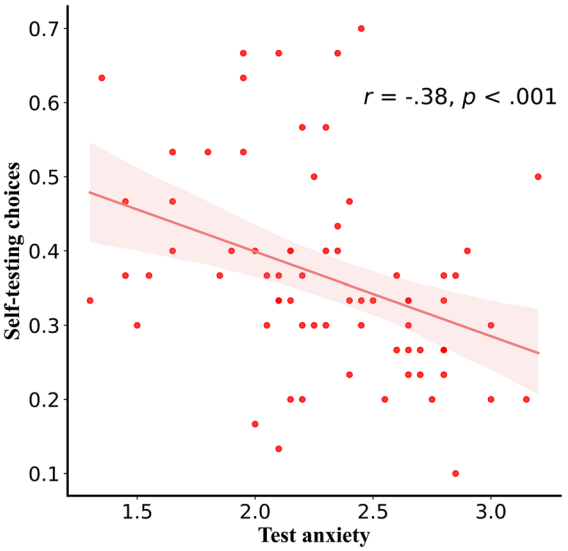

Fig. 3 Relationship between self-testing choices and TA in Experiment 2

Table 2 Regression model for self-testing choices in Experiment 2

Model		*B*	*S.E.*	*β*	*t*	*p*	95% CI of *B*
H_0	(Intercept)	0.06	0.01		0.57	0.57	[−0.14, 0.25]
	FTP	0.33	0.12	0.31	2.73	0.008	[0.09, 0.57]
	EB	0.02	0.02	0.13	1.07	0.29	[−0.02, 0.06]
	ToI	0.03	0.02	0.17	1.39	0.17	[−0.01, 0.08]
Model summary of H_0: $F(3, 67) = 4.01$, $p = .01$, $R^2 = 0.15$							
H_1	(Intercept)	0.33	0.15		2.24	0.03	[0.04, 0.63]
	FTP	0.25	0.12	0.24	2.09	0.04	[0.01, 0.50]
	EB	0.02	0.02	0.13	1.17	0.25	[−0.02, 0.06]
	ToI	0.02	0.02	0.08	0.66	0.51	[−0.03, 0.07]
	TA	−0.09	0.04	−0.28	−2.41	0.02	[−0.16, −0.02]
Model summary of H_1: $F(4, 66) = 4.67$, $p = .002$, $R^2 = .22$							
Model comparison between H_0 and H_1: $\Delta F(1, 66) = 5.80$, $p = .02$, $\Delta R^2 = .07$							

FTP = final test performance; *EB* = effort belief; *ToI* = theory of intelligence; *TA* = test anxiety; *S.E.* = standard error

Note that although participants made strategy choices during the initial study phase, they did not actually restudy or take a test on any items before the final test, for reasons explained above.

Results

Table 1 shows the correlation matrix of measures collected in this experiment. Significant correlations were found between TA and theory of

intelligence ($r = 0.29$, $p = 0.01$, $BF_{10} = 2.81$), between effort belief and theory of intelligence ($r = 0.33$, $p = 0.005$, $BF_{10} = 7.12$), and between self-testing choices (i.e., proportions of image-name pairs selected to be tested) and final test performance ($r = 0.31$, $p = 0.009$, $BF_{10} = 4.15$). Of most importance, there was a negative correlation between TA and self-testing choices, $r = -0.38$, $p = 0.001$, $BF_{10} = 24.63$ (see Fig. 3), conceptually replicating the main finding of Experiment 1.

A hierarchical multiple linear regression analysis was conducted to quantify the effect of TA on self-testing choices, with effort belief, theory of intelligence, and final test performance (an index of memory ability) as control variables (see Table 2 for detailed results). Specifically, effort belief, theory of intelligence, and final test performance were included in level 1 (model H_0), and TA was included in level 2 (model H_1). Overall, the results indicated that the first (H_0) model was significant, $F(3, 67) = 4.01$, $p = 0.011$, $R^2 = 0.15$. Only final test performance positively predicted self-testing choices, $b = 0.33$ [0.09, 0.57], $t = 2.73$, $p = 0.008$, suggesting that, as predicted, memory ability affects self-testing choices. The second (H_1) model was also significant, $F(4, 66) = 4.67$, $p = 0.002$, $R^2 = 0.22$. Most importantly, including TA as a predictor significantly improved the goodness of model fit, $\Delta F(1, 66) = 5.80$, $p = 0.02$, $\Delta R^2 = 0.07$. In the second model, both final test performance, $b = 0.25$ [0.01, 0.50], $t = 2.09$, $p = 0.04$, and TA, $b = -0.09$ [−0.16, −0.02], $t = 2.41$, $p = 0.02$, significantly predicted the proportion of items selected to be tested.

Discussion

In Experiment 2, TA negatively predicted self-testing choices in a self-regulated learning task, replicating the main finding of Experiment 1. Additionally, Experiment 2 further demonstrated that TA negatively predicted self-testing choices even when the confounding effects of effort belief, theory of intelligence, and memory ability were controlled for.

Experiment 3

Experiments 1 and 2 showed that individuals with high TA are less likely to test themselves. Experiment 3 asks whether learners would avoid self-testing when preparing for a high-stake assessment. As proposed by the transaction model of stress, cognitive TA (that is, one's evaluation and appraisal of a test event) can affect individuals' beliefs and behaviors outside the test contexts (Spielberger & Vagg, 1995). Additionally, the TAM hypothesis claims that learners typically regard testing as a monitoring (rather than learning) tool. Accordingly, we predict that differential evaluations of high- and low-stake tests would then affect learners' self-testing choices. To test this prediction, a quasi-experiment (Experiment 3) was conducted in which we directly manipulated the stake level of a middle-term exam in a Statistics for Psychology course.

Participants and Experimental Design

Twenty-five undergraduate students, who took the Statistics for Psychology course taught by the last author, participated in this experiment. All participants were sophomores. They gave informed consent to participate and for their data to be used for research purposes. Because the total number of students in this course was relatively small, we manipulated test stake (high vs. low) by using a within-subjects design (see below for details). Another reason for employing a within-subjects design was to avoid potential confounding effects of other individual differences variables, such as effort belief, theory of intelligence, and learning ability. We acknowledge that the sample size in this experiment was determined by the class size and was not pre-determined.

Materials and Procedure

The stimuli were 20 statistics topics (e.g., *F*-distribution, one-way ANOVA, repeated measures ANOVA, simple-effect analysis, post hoc analysis, Pearson's *r* correlation, and Spearman's rank correlation). These topics were divided into two sets, pre-determined by the course instructor, with 10 topics in each set. Importantly, list assignment to the high- and low-stake conditions was counterbalanced across participants.

The experiment took place in a class 1 week before the course mid-term exam. During the class, the instructor informed students that he had published practice test questions and knowledge summaries on an online platform named StatsLearning, and they could use this platform to prepare for the mid-term exam. On the platform, there were practice test questions and knowledge summaries relating to 20 statistics topics that might be assessed in the mid-term exam. For each topic, students made a choice about whether they would like to review a summary of key knowledge points on that topic or take a practice test on it. If they chose "restudying" (i.e., reviewing), they reviewed a summary of the selected topic on the platform. If they selected testing, the platform provided them with practice test questions and corrective feedback for the topic. Additionally, students were explicitly told that the two learning formats did not differ in how much they covered a given topic. Finally, but importantly, they were informed that their mid-term exam scores on 10 topics (i.e., high-stake topics) would contribute to 30% of their final course grade and their scores on the other 10 topics (i.e., low-stake topics) would not contribute. They would see whether mid-term exam performance on a given topic would contribute to the final course grade in the following survey.

Next, all students received an electronic survey, programmed via Credamo (a platform specialized for questionnaires and online behavioral experiments), and were asked to complete the survey on their mobile phone. The 20 topic names were presented one by one in random order. Below each topic name, a sentence informed students whether their mid-term exam scores on that topic would contribute to their final course grade. They were also provided with two options (restudy vs. test) to

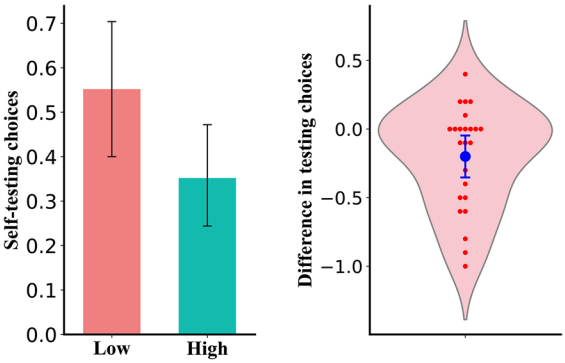

Fig. 4 Self-testing choices as a function of test stake in Experiment 3. Note: In the violin plot, each red dot represents the difference in self-testing choices between the high- and low-stake conditions for one participant, with the blue point representing the group average of the difference scores. Error bars indicate 95% CI

report which strategy they would like to use on the platform to prepare for the mid-term exam. There was no time pressure for making a strategy choice.

After all students completed the questionnaire, they were informed that the stage they had just completed was for research purposes and that practice test scores would not contribute to the mid-term exam scores. Finally, they were debriefed.

Results

Collapsing across the high- and low-stake conditions, the proportion of self-testing choices ($M = 0.45$, $SD = 0.29$), calculated as the proportion of topics selected to be tested, was not statistically different from the chance level (i.e., 0.50), difference = −0.05, 95% CI [−0.17, 0.07], $t(24) = -0.82$, $p = 0.42$, Cohen's $d = -0.16$, $BF_{10} = 0.29$. Critically, as shown in Fig. 4, students chose self-testing less frequently in the high-stake condition ($M = 0.35$, $SD = 0.31$) than in the low-stake condition ($M = 0.55$, $SD = 0.38$), difference = −0.20, 95% CI [−0.35, −0.05], $t(24) = -2.71$, $p = 0.01$, $d = -0.54$, $BF_{10} = 4.00$, suggesting that students are less likely to test themselves when preparing for high- than for low-stake exams. Among the 25 students, 13 selected self-testing less frequently in the high-stake condition than in the low-stake condition, with only 5 showing the converse pattern (there were 7 ties).

In the low-stake condition, self-testing choices ($M = 0.55$, $SD = 0.38$) were close to the chance level, difference = 0.05, 95% CI [−0.10, 0.21], $t(24) = 0.69$, $p = 0.50$, Cohen's $d = 0.14$, $BF_{10} = 0.26$, indicating no preference between testing and restudying when preparing for a low-stake exam. However, in the high-stake condition, self-testing choices ($M = 0.35$, $SD = 0.31$) were significantly lower than the chance level, difference = −0.15, 95% CI [−0.28, −0.02], $t(23) = -2.39$, $p = 0.03$, Cohen's $d = -0.48$, $BF_{10} = 2.23$, indicating that students overwhelmingly prefer restudying over testing when preparing for a high-stake exam.

Discussion

Experiment 3 reveals that a high-stake exam decreases students' self-testing preference in applied educational settings. It is worth noting that the experiment employed a within-subjects block design, in which two blocks of statistical concepts were

assigned to the high-stake condition and two to the low-stake condition. In such a design, it is unlikely that students' state anxiety (i.e., concurrent feelings of worry) would frequently increase and decrease, even though it has been well established that high-stake assessments are more anxiety-provoking than low-stake ones (Endler & Kocovski, 2001; Wenzel & Reinhard, 2021). Hence, it is unlikely that state anxiety is the main driver of the test stake effect on self-testing choices observed here. By contrast, students' cognitive evaluation and appraisal of the stake level of different concepts (i.e., cognitive TA) is a more plausible factor.

Specifically, according to the transactive model of stress, students cognitively worry about poor exam performance relating to high-stake concepts. Additionally, the TAM hypothesis proposes that learners typically regard self-testing as an opportunity for diagnosing their learning status and consider restudying as an opportunity for further improving learning (Badali et al., 2022). In Experiment 3, students might have perceived low-stake concepts as relatively less important (i.e., less stressful) and hence chose more items to be tested with the aim to check their knowledge mastery level. By contrast, they worried about poor exam performance regarding high-stake concepts and hence chose more concepts to restudy in the hope of achieving superior exam performance.

Overall, the test stake effect on self-testing choices observed in Experiment 3 is unlikely to be attributable to differences in state anxiety between the high- and low-stake conditions. By contrast, the transactive model of stress and the TAM hypothesis provide a coherent explanation: Cognitive TA (i.e., one's cognitive evaluation and appraisal of test stake) and beliefs about self-testing and restudying (i.e., beliefs that self-testing is an opportunity for checking learning status, and restudying is a tool for improving learning) jointly contribute to the test stake effect on self-testing choices.

Experiment 4

Experiment 4 had two aims. The first was to conceptually replicate the main findings of Experiment 3 with a different manipulation of test stake. Experiment 4 adopted Wenzel and Reinhard's (2021) method in which manipulation of test stake was achieved by informing participants whether their test scores would be released to their groupmates, a scenario simulating real classroom situations. The second aim of Experiment 4 was to examine whether avoidance of self-testing induced by a high-stake test would lead to poor learning performance.

Participants and Design

Like Experiment 3, Experiment 4 adopted a within-subjects design (stake level: high vs. low) to avoid potential confounding effects of other individual differences variables. The sample size was determined according to the effect size of the test stake on

self-testing choices observed in Experiment 3 ($d = -0.54$). To achieve statistical power of at least 0.80 for a paired comparison, the required sample size was 29.

Due to over-recruitment, we finally collected data from 36 undergraduates (26 female; $M_{age} = 20.00$, $SD = 1.74$) from the participant pool at Beijing Normal University. They completed the task in groups in a classroom, with 6 participants randomly assigned to each group. They confirmed that they were not Biology, Psychology, or Neuroscience majors, gave informed consent, and received monetary compensation.

Materials

The learning materials were 40 image-name pairs of human anatomical structures, which were divided into four lists, with 10 pairs in each list. List difficulty was matched according to the pilot study described in the "Method" section of Experiment 2. According to the pilot data, there were no differences in item difficulty among the four lists, $F(3, 36) < 0.001$, $p = 1.00$. Two lists were randomly assigned to the low-stake condition, with the other two assigned to the high-stake condition. List assignment to conditions was counterbalanced across participants.

Procedure

Participants completed the task in small groups of 6 in a classroom. They sat around a table and used laptops to complete the experiment. Participants were informed that they would study four lists of anatomical structures in preparation for a final memory test. After completing the final test, scores on two lists would be released to their groupmates (high-stake lists), while scores for the other two lists would not be released and kept private (low-stake lists). Before learning each list, the computer would inform them whether their final test performance (rather than practice test performance) of that list would be released to groupmates or not. After the instructions, participants practiced three items in order to understand the experimental procedure.

The experiment consisted of three stages: initial study, review, and final test. In the initial study phase, participants studied four lists of anatomical structures one by one and list by list. High- and low-stake lists were alternated. Whether participants started with a high- or low-stake list was counterbalanced.

The initial study phase was similar to that of Experiment 2. Before studying each list, the computer informed participants whether the final test performance of the next list would be released to groupmates. Next, the 10 image-name pairs were presented one by one in random order. Each item was presented for 10 s, and after studying it, participants selected which strategy (restudy vs. test) they would like to use to review the item in the review phase. After initial learning of all four lists, participants completed a distractor task identical to that of Experiment 2. Then, the review phase started, during which the 40 items were presented one by one in random order, with each item reviewed in a format that honored participants' strategy choices. That is, for each restudy trial, the structure image and its name were simultaneously shown on the screen for 10 s for participants to restudy. By contrast, for

each test trial, the structure image was shown on screen for 7 s with a blank box shown below it, and participants inputted the corresponding structure name into the blank box, after which corrective feedback (i.e., image accompanied by structure name) was provided on screen for 3 s.

After the review phase, participants completed the same distractor task (but with new mathematical problems) for 2 min. Finally, they took a final cued recall test, identical to that in Experiment 2. At the end of the experiment, participants were debriefed and informed that in fact none of their test scores would be released to their groupmates.

Results

Across the high- and low-stake conditions, self-testing ($M = 0.36$, $SD = 0.17$) was selected significantly less frequently than the chance level (i.e., 0.50), difference $= -0.14$, 95% CI $[-0.20, -0.09]$, $t(35) = -5.14$, $p < 0.001$, $d = -0.86$, $BF_{10} > 1000$, indicating that participants overwhelmingly preferred restudying over self-testing. As shown in Fig. 5a, self-testing choices were significantly lower in the high-stake condition ($M = 0.28$, $SD = 0.17$) than in the low-stake condition ($M = 0.43$, $SD = 0.23$), difference $= -0.15$, 95% CI $[-0.23, -0.07]$, $t(35) = -4.00$, $p < 0.001$, $d = -0.67$, $BF_{10} = 88.47$, re-confirming that learners are less likely to test themselves when preparing for high-stake tests. Twenty-six

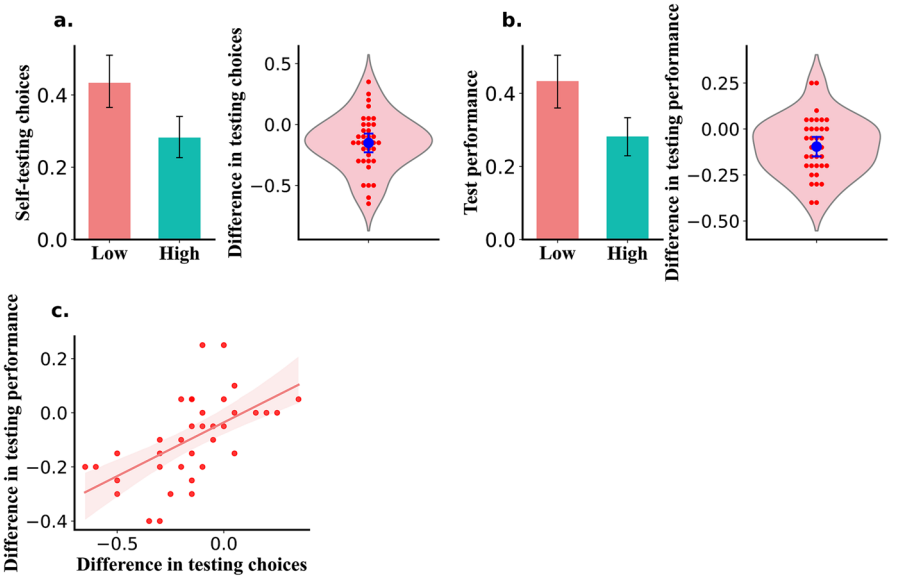

Fig. 5 Self-testing choices and final test performance in Experiment 4. Note: Self-testing choices (**a**) and final test performance (**b**) in the high- and low-stake conditions. Correlation between the difference in self-testing choices and the difference in final test performance (**c**). In the violin plots, each red dot represents the difference in self-testing choices (**a**) and final test performance (**b**) for one participant, with the blue points representing the group average difference scores. Error bars indicate 95% CI

participants selected fewer items for self-testing in the high-stake condition than in the low-stake condition, with only 7 showing the converse pattern (there were 3 ties). Importantly, as shown in Fig. 5b, test performance was significantly poorer in the high-stake condition ($M = 0.47$, $SD = 0.13$) than in the low-stake condition ($M = 0.37$, $SD = 0.09$), difference $= -0.10$, 95% CI $[-0.15, -0.04]$, $t(35) = -3.66$, $p < 0.001$, $d = -0.61$, $BF_{10} = 37.79$. Twenty-four participants performed worse in the high-stake condition than in the low-stake condition, and 8 showed the reverse pattern (there were 5 ties).

Lastly, a within-subjects mediation analysis was conducted via the *MEMORE* package in SPSS 25.0 (Montoya, 2019). This analysis examined whether self-testing usage mediated the negative effect of test stake on test performance. As shown in Fig. 6, the direct effect of test stake on final test performance was not significant, $c' = -0.04$, 95% CI $[-0.09, 0.02]$. Critically, a bias-corrected bootstrap resampling analysis (with 5000 resamples) and normal theory tests (Fairchild & MacKinnon, 2008) indicated that the mediating effect of self-testing choices was significant, $a*b = -0.06$, 95% CI $[-0.10, -0.03]$. These findings suggest that the effect of test stake on final test performance was completely mediated by its effect on self-testing usage. Consistent with the mediation results, the differences in self-testing choices between the high- and low-stake conditions significantly predicted the differences in final test performance, $r = 0.57$, $p < 0.001$, $BF_{10} = 125.10$ (see Fig. 5c).

Discussion

Experiment 4 confirmed that learners are less likely to test themselves when preparing for high-stake tests, and self-testing avoidance impairs learning. The effect of test stake on learning was fully mediated by its indirect effect via self-testing usage. As discussed above, the transactive model of stress and the TAM hypothesis jointly explain the test stake effect on self-testing choices.

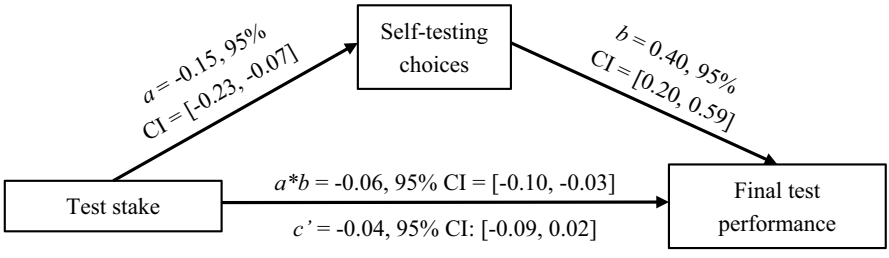

Fig. 6 Mediation results in Experiment 4

Experiment 5

Experiments 3 and 4 demonstrated that participants are less likely to choose self-testing when confronting a high-stake test, which consequently impairs learning (Experiment 4). A possible explanation for the test stake effect on self-testing usage observed in those experiments is that cognitive TA (i.e., one's cognitive evaluation of the stake level of an upcoming assessment) and beliefs about testing and restudying (i.e., beliefs that self-testing is a tool for diagnosing learning status while restudying is more effective for improving learning) lead to avoidance of self-testing induced by high-stake tests. However, it should be noted that neither Experiment 3 nor 4 directly measured state anxiety in the high- and low-stake conditions. Hence, it remains unknown whether state anxiety affects self-testing usage and, for this reason, impairs learning. Experiment 5 was conducted to explore these two questions. Specifically, Experiment 5 employed a between-subjects manipulation of test stake and included a state anxiety manipulation check in both the high- and low-stake conditions.

Participants

One hundred and fourteen undergraduate students ($M_{age} = 18.87$, $SD = 0.61$; 64 females) were recruited from two teaching classes at Tianjin Vocational Institute, with 57 students in each class. The sample size was determined by the class size. One class was randomly assigned to the high-stake condition and the other to the low-stake condition. All students in each class completed the task together in a classroom equipped with 70 computers. They provided informed consent and received monetary compensation.

Materials and Procedure

The materials were identical to those in Experiment 2. In addition to trait TA measured by the TAI at the beginning of the experiment (i.e., before the stake manipulation, Cronbach's $\alpha = 0.74$), students' state anxiety before and after the test stake manipulation phase was also measured. State anxiety was measured by the 20-item version of the State-Trait Anxiety Inventory (STAI-S; Spielberger et al., 1983). Participants rated each item describing their concurrent feelings (e.g., "*I feel calm*") from 1 (*not at all*) to 4 (*very much*). Responses to 10 negative items were reverse-scored. The average rating across the 20 items was taken as an index of state anxiety. The Cronbach's α for the three measures of state anxiety of the current sample was 0.67, 0.71, and 0.65.

The measurements of state anxiety were taken three times across the experiment: at the beginning of the experiment, immediately after the test stake manipulation phase, and immediately after the initial study phase. The first measurement provided a baseline, and the average of the second and third measurements was taken as an index of state anxiety experienced during the initial study phase. We hypothesized

that if the test stake manipulation was successful, the high-stake group would experience higher levels of state anxiety (represented by the average of state anxiety reported at the second and third measurements) than the low-stake group.

All students undertook the experiment on an individual computer in a classroom. The procedure was similar to that of Experiment 4, consisting of three phases, including initial study, review, and final test. At the beginning of the experiment, students in both conditions completed the TAI and STAI-S scales (i.e., the first measurement of state anxiety). Then, for the class in the high-stake condition, students were informed that immediately after the final test, their final test scores and rank order would be displayed on a PowerPoint slide in the classroom, and all of their classmates would see their scores and rank. To enhance the credibility of the instructions, the experimenter showed an example of a performance-and-ranking table to participants in the high-stake condition. By contrast, for the class in the low-stake condition, students were informed that their final test scores would be anonymous and kept confidential, and they would receive a mini-lecture on educational psychology after finishing the final test.

After receiving the instructions, all students in both conditions completed the STAI-S scale again (the second measurement of state anxiety). Then, the initial study session commenced in which all students learned 30 anatomical structures in random order. After studying each structure, they made a strategy choice between "restudy" and "test." After studying all structures, participants in both conditions again completed the STAI-S scale (the third measurement of state anxiety). Next, all participants solved mathematics problems for 2 min. Then, they reviewed all items in a manner honoring their study strategy choices. After the review session, they completed a new set of mathematics problems for another 2 min. Finally, they completed the final cued recall test, identical to that in Experiment 4. At the end of the experiment, all participants were debriefed and were informed that their test scores would not be released to their classmates.

Results

There was no difference in trait TA between the low- ($M = 2.14$, $SD = 0.52$) and high-stake ($M = 2.15$, $SD = 0.59$) groups, difference = −0.01, 95% CI [−0.21, 0.20], $t(112) = 0.07$, $p = 0.95$, Cohen's $d = 0.01$, $BF_{10} = 0.20$, suggesting no baseline difference in trait TA between the two groups.

Frequentist and Bayesian mixed analyses of variances (ANOVAs) were conducted on state anxiety, with group as a between-subjects factor and measurement time (before vs. after the test stake manipulation) as a within-subjects factor (Fig. 7a). Specifically, for each participant, we calculated an average of state anxiety reported at the second and third measurements and took this composite score as an index of state anxiety measured after the test stake manipulation phase. The ANOVA revealed a main effect of group, $F(1, 112) = 14.52$, $p < 0.001$, $\eta_p^2 = 0.12$, $BF_{10} = 77.18$, with higher state anxiety in the high-stake condition than in the low-stake condition (see Fig. 7b), indicating that the test stake manipulation was successful. There was also a main effect of measurement time, $F(1, 112) = 63.87$,

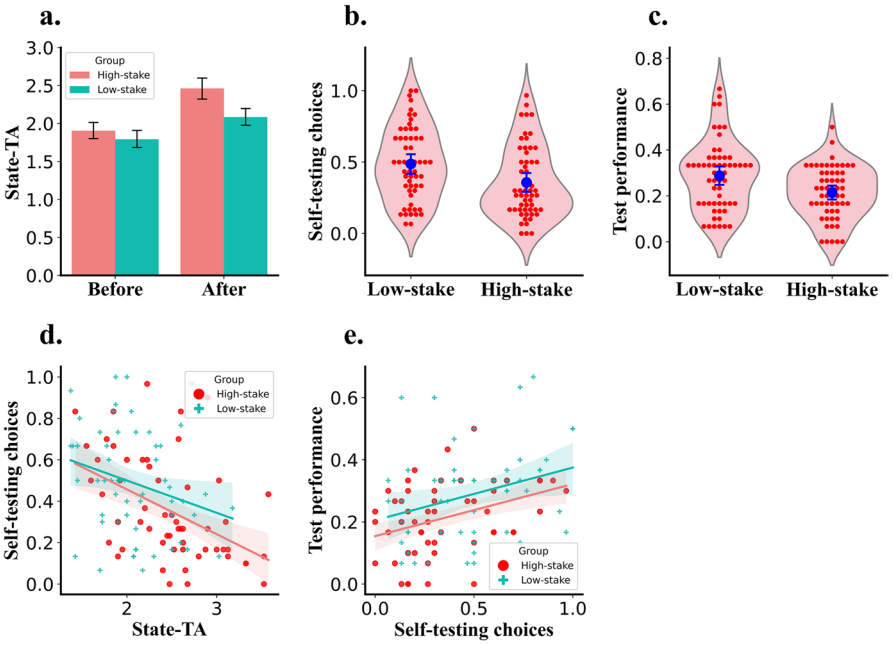

Fig. 7 Results of Experiment 5. Note: (**a**) state anxiety measured before and after the test stake manipulation. Self-testing choices (**b**) and final test performance (**c**) in the high- and low-stake conditions. Relationship between self-testing choices and state anxiety (**d**) and between self-testing choices and final test performance (**e**). In the violin plots, each red dot represents one participant's data, with the blue points representing group averages. Error bars indicate 95% CI

$p < 0.001$, $\eta_p^2 = 0.36$, $BF_{10} > 1000$, with higher state anxiety reported after than before the test stake manipulation.

Critically, the interaction between group and measurement time was also significant, $F(1, 112) = 6.20$, $p = 0.014$, $\eta_p^2 = 0.05$, $BF_{10} = 3.51$. Further tests showed that the high- ($M = 1.91$, $SD = 0.40$) and low-stake ($M = 1.79$, $SD = 0.40$) groups differed minimally before the manipulation, difference = 0.11, 95% CI [−0.04, 0.26], $t(112) = 1.50$, $p = 0.14$, Cohen's $d = 0.28$, $BF_{10} = 0.54$. However, after the manipulation (i.e., during the initial study phase), state anxiety was significantly higher in the high-stake group (2.46, $SD = 0.53$) than in the low-stake ($M = 2.09$, $SD = 0.43$) group, difference = 0.38, 95% CI [0.20, 0.57], $t(112) = 4.16$, $p < 0.001$, Cohen's $d = 0.78$, $BF_{10} = 325.76$, again confirming that the test stake manipulation was successful.

Of most importance, as shown in Fig. 7b, self-testing choices were significantly lower in the high-stake group ($M = 0.36$, $SD = 0.25$) than in the low-stake ($M = 0.49$, $SD = 0.26$) group, difference = −0.13, 95% CI [−0.22, −0.04], $t(112) = 2.73$, $p = 0.007$, Cohen's $d = 0.51$, $BF_{10} = 5.28$, replicating the test stake effect on self-testing choices observed in Experiments 3 and 4. Specifically, the frequency of self-testing choices did not differ from chance (i.e., 50%) in the low-stake group, difference = −0.01, 95% CI [−0.08, 0.05], $t(56) = -0.39$,

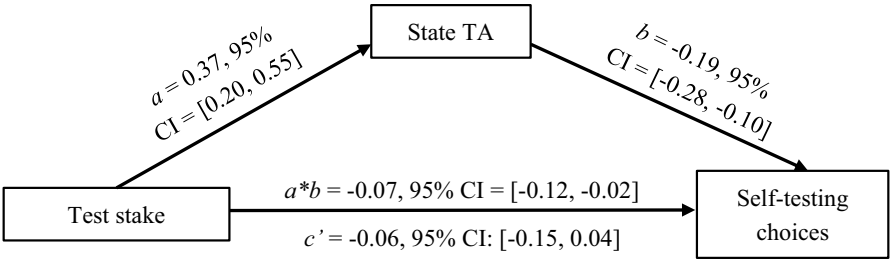

Fig. 8 Mediation results of Experiment 5

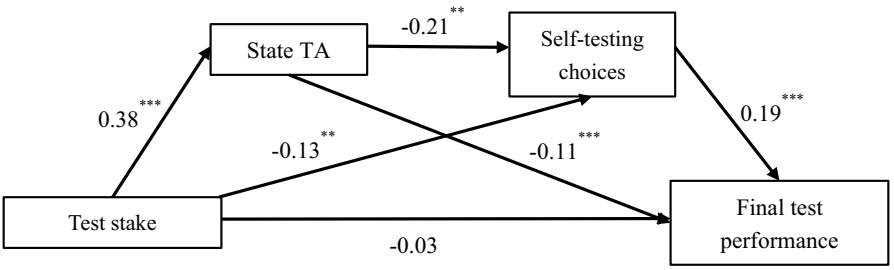

Fig. 9 Chain mediation results of Experiment 5

$p = 0.70$, Cohen's $d = -0.05$, $BF_{10} = 0.16$, but did in the high-stake group, difference $= -0.14$, 95% CI $[-0.21, -0.08]$, $t(56) = -4.33$, $p < 0.001$, Cohen's $d = -0.57$, $BF_{10} = 345.98$. Furthermore, as shown in Fig. 7c, recall performance was poorer in the high-stake group ($M = 0.21$, $SD = 0.11$) than in the low-stake group ($M = 0.29$, $SD = 0.15$), difference $= -0.07$, 95% CI $[-0.12, -0.02]$, $t(112) = -2.94$, $p = 0.004$, Cohen's $d = -0.55$, $BF_{10} = 8.82$. These results successfully replicated the negative effect of the test stake on learning performance observed in Experiments 3 and 4.

Next, a mediation analysis was conducted to check the indirect effect of group (i.e., test stake) on self-testing choices via state anxiety (i.e., the average of state anxiety reported at the second and third measurements). As shown in Fig. 8, the direct effect was not significant, $c' = -0.06$, 95% CI $[-0.15, 0.04]$. Importantly, the indirect effect of the group on self-testing choices via state anxiety was significant, $a*b = -0.07$, 95% CI $[-0.12, -0.02]$. These results suggest that state anxiety completely mediated the effect of test stake on self-testing choices.

Finally, a chain mediation analysis was conducted via the SPSS macro *PROCESS Model 6* (Abu-Bader & Jones, 2021), with 5000 resamples to quantify the chain mediation effect of the group on final test performance via state anxiety and self-testing choices. As shown in Fig. 9, the direct effect of group (i.e., test stake) on final test performance was not significant, $\beta = -0.030$, 95% CI $[-0.078, 0.020]$. The indirect effect of group on final test performance via state anxiety was significant, $\beta = -0.028$, 95% CI $[-0.050, -0.080]$. The indirect effect of group on final test performance via self-testing choices was not significant, $\beta = -0.008$, 95% CI $[-0.025,$

0.040]. Most importantly, the chain mediation effect (that is, group → state anxiety → self-testing choices → final test performance) was significant, $\beta = -0.010$, 95% CI [−0.022, −0.002]. Figures 7d and 7e visualize the relationships from state anxiety to self-testing choices, $r = -0.42$, $p < 0.001$, $BF_{10} > 1000$, and self-testing choices to final test performance, $r = 0.36$, $p < 0.001$, $BF_{10} = 420.07$.

Discussion

Experiment 5 confirmed that the high-stake test provoked state anxiety and decreased self-testing choices. More importantly, avoidance of self-testing subsequently produced an impairment in final test performance. Notably, there was no baseline difference in trait TA between the two groups. This means that, regardless of individual differences in trait TA, high-stake tests can generate a detrimental effect on students' self-testing choices by provoking high state anxiety.

General Discussion

Across five experiments, the present study consistently documented a negative correlation between TA (or test stake) and self-testing usage. In particular, this relationship was confirmed by a survey study (Experiment 1), two quasi-experiments (Experiments 3 and 5), and two laboratory experiments (Experiments 2 and 4). Additionally, Experiments 1, 4, and 5 demonstrated that avoidance of self-testing caused by high TA or high test stake impaired learning. Although existing evidence has provided suggestive findings about a negative association between TA and self-testing usage by showing that students are more likely to implement self-testing when good test performance is expected (e.g., when the test is expected to be relatively easy to perform; Rivers, 2021), the present study is the first to directly and empirically confirm a negative effect of TA (or test stake) on self-testing usage.

There are several explanations for why individual differences in TA predict individual differences in self-testing usage, as demonstrated in Experiments 1 and 2. The first concerns uncomfortable feelings (e.g., extensive worry about retrieval failure) associated with TA. During the test, individuals with high TA may experience heightened levels of anxiety, worry, and pathological physiological arousal, even when the test is low stake (Roos et al., 2021). To avoid these uncomfortable feelings, students may choose to use "safe" strategies (e.g., restudying) and avoid testing themselves during learning. Aside from avoidance of negative feelings aroused by tests, previous studies suggested that individuals with high TA hold an "offline" aversion toward testing during the test preparation phase (Cassady, 2004). That is, they typically view testing as a more threatening strategy, relative to individuals with low TA. Additionally, individuals with high TA may be more sensitive to negative feedback (e.g., retrieval failure or poor test performance), which will also reduce their willingness to implement self-testing (Clark & Svinicki, 2015; Morris & Fulmer, 1976; Vaughn & Kornell, 2019).

Another possible explanation is that students with high TA may not appreciate the benefits of testing; namely, they may lack metacognitive awareness that testing is a powerful strategy for enhancing learning. Numerous studies have confirmed that metacognitive beliefs about strategy effectiveness affect strategy usage (Hui et al., 2021; Sun et al., 2022). Lack of awareness of test-enhanced learning may be another reason for underemployment of self-testing, especially for individuals with high TA. One of our recent studies observed supportive findings for this explanation (Liu et al., 2023). In this study, participants were asked to read two scientific passages and then imagined that they would use one of two strategies to review the two passages: (a) restudying them twice and (b) taking a free recall test twice. Next, they were asked to predict how much information they would be able to recall from the passages on a final test administered 1 week later, if they used the test (or restudy) strategy to review the passages. They made their predictions (i.e., JOLs) on a scale ranging from 0 (*not at all*) to 100 (*all of the information*). The results showed that participants provided overall higher JOLs for tested passages than for restudied ones, suggesting that learners do metacognitively appreciate the benefits of testing (for related findings, see Weissgerber & Rummer, 2023). More importantly, the results revealed that participants' trait TA negatively predicted their metacognitive awareness of test-enhanced learning (represented as the signed difference in JOLs between the test and restudy conditions), suggesting that the higher their level of trait TA, the poorer their metacognitive awareness about the benefits of test-enhanced learning. Hence, it is reasonable to hypothesize that high-TA learners are less likely to appreciate the merits of testing, and this metacognitive unawareness prevents them from engaging in self-testing during self-regulated learning (Bjork et al., 2013; Rivers et al., 2022; Roediger & Karpicke, 2006).

Experiments 3 and 4 demonstrated that participants were less inclined to self-test when preparing for high-stake tests (or exams). This suggests that they may regulate their self-testing choices according to their cognitive evaluation of the potential consequences of poor performance on high- and low-stake contents. This cognitive TA could also act as a mechanism underlying the effect of test stake on self-testing choices observed in these two experiments. According to transactional models (Spielberger & Vagg, 1995), individuals' responses to a stressful situation rely upon their situational appraisal. Furthermore, as postulated by the TAM hypothesis, learners frequently regard testing as a monitoring tool and consider restudying as a learning tool (Badali et al., 2022), consequently believing that prioritizing restudying over testing will more reliably improve their performance and help alleviate their worries about performing poorly on high-stake contents (Rivers, 2021).

Different from Experiments 3 and 4, Experiment 5 adopted a between-subjects manipulation of test stake, aiming to investigate whether high state anxiety, provoked by high test stake, affects self-testing usage and learning. In line with prior research (Hinze & Rapp, 2014; Wenzel & Reinhard, 2021), the test stake manipulation was successful as participants in the high-stake group reported higher levels of state anxiety compared to those in the low-stake group. Crucially, high state anxiety reduced self-testing and led to poorer learning. Therefore, in addition to cognitive evaluations of the test situations (i.e., cognitive TA), high state anxiety provoked by the high test stake can also engender avoidance of self-testing. A reasonable

conjecture for the observed negative impact of state anxiety on self-testing choices is that high state anxiety triggers general aversion toward testing (Clark & Svinicki, 2015; Vaughn & Kornell, 2019), even though the stake level of self-tests (i.e., practice tests) is low.

The negative effect of TA (or test stake) on self-testing usage observed here also echoes the divergent findings regarding self-testing choices observed between lab-based and online-based studies. For instance, a recent online study used similar structure-name pairs as in the present study, in which participants made item-by-item choices of testing vs. restudying (Fan et al., 2023). Strikingly, and unlike the present study, Fan et al.'s experiments showed that participants overwhelmingly preferred testing over restudying. The proportions of items selected to be tested were 68%, 72%, and 65% in their Experiments 1–3, respectively. Another online experiment by Tullis et al. (2018) also assessed self-testing preference in an online experiment and observed that participants chose 63% of items to be tested. By contrast, in the present Experiments 2–5, the proportions of study items selected to be tested were either significantly or numerically lower than the chance level (i.e., 50%). A possible explanation for these divergent findings is that lab-based experiments trigger higher levels of state TA than online-based experiments because in the laboratory participants complete the learning task under the experimenters' face-to-face supervision, and as stated by social derogation theory (Branscombe & Wann, 1994; Symes & Putwain, 2020), laboratory participants may worry that poor performance would affect their self-interests (e.g., self-esteem). Although there is no direct evidence showing that participants are more anxious during lab-based than online experiments, there are suggestive data that students experienced lower levels of TA during online compared with offline learning and testing (Stowell & Bennett, 2010). Particularly, switching to online learning has been proposed as an explanation for why students' TA was reduced during COVID-19 (Ewell et al., 2022). Future studies are encouraged to experimentally test whether students are more likely to implement self-testing during online than offline learning and whether this difference is caused by variance in TA.

Previous studies have mostly attributed underemployment of self-testing to a lack of metacognitive awareness of the benefits of test-enhanced learning (for a review, see Rivers, 2021). The present study revealed that erroneous metacognitive beliefs are not the only factor constraining self-testing usage. Students may decline to self-test due to anxious feelings induced by tests, especially for individuals with high trait TA and when they are encountering a high-stake test. Our findings are also interpretable by and consistent with the KBCP framework (McDaniel & Einstein, 2020): Learners are reluctant to take tests because of high TA or high-stake tests (insufficient commitment), even when they explicitly know what self-testing is and how to implement it (knowledge), appreciate the benefits of self-testing (belief), and have set up a plan of implementing self-testing (plan). In particular, in Experiment 3, even when we presumed that students might use the learning platform to test themselves as a mock before the mid-term exam, the results showed the exact reverse pattern: Students preferred restudying over self-testing in the high-stake condition. In fact, recent evidence has shown that learners do not always underestimate

the effectiveness of testing (Rea et al., 2022; Weissgerber & Rummer, 2023). Thus, TA (both emotional and cognitive TA) may be a critical factor that undermines the commitment to self-testing regardless of whether learners appreciate its benefits.

In a recent review, McDaniel and Einstein (2020) proposed several approaches to intervene in students' study strategies. For example, students' tendency to use a strategy can be increased by a *utility-value intervention* that addresses the (intrinsic or extrinsic) reward of using the strategy (e.g., "*This strategy will help me get accepted into my dream college*"; Johnson & Sinatra, 2013). Alternatively, a *growth mindset intervention* can boost the usage of an effortful study strategy by encouraging students to believe that their ability is incremental. With a growth mindset, students are more likely to develop their learning strategies (Yeager, et al., 2016). However, as shown in the present study, theory of intelligence did not predict self-testing choices. Although multiple interventions have been proposed to change students' study strategies, these approaches have not yet been examined with regard to self-testing. Among all learning strategies, self-testing is unusual as it is normally accompanied by anxiety and negative feedback. Thus, we are skeptical about the effectiveness of these interventions for individuals with high TA, as these approaches do not take into account students' fear of failure and feelings of unease when taking tests. The present study highlights the crucial importance of TA interventions in promoting self-testing usage (for reviews of TA interventions, see Ergene, 2003; Soares & Woods, 2020; von der Embse et al., 2018).

The negative association between TA (or test stake) and self-testing provides an alternative explanation for why TA causes learning deficits. Specifically, students with high TA are reluctant to test themselves in daily learning settings. This not only impairs their learning but also enlarges the academic achievement gap among students with different levels of TA (Yang et al., 2021). This account does not place emphasis on how TA might interfere with cognitive functions during the test phase. Instead, we argue that avoidance of utilizing effective learning strategies (i.e., self-testing) is a critical factor in explaining the long-term learning deficits induced by high TA. Although the present study does not challenge the possibility of non-interference mechanisms, we provided evidence and a potential explanation for Theobald et al.'s (2022) findings that it is how well students master knowledge during the exam preparation phase rather than how anxious the students are during the test phase that accounts for poor exam performance caused by high TA.

Besides these theoretical implications, the findings also bear practical implications for education. An important goal of previous research and systematic reviews on the testing effect is to encourage students to actively use self-testing to boost learning (e.g., Hartwig & Dunlosky, 2012; Kornell & Son, 2009; Yang et al., 2021). Previous studies showed that such a powerful strategy is generally underemployed in real educational settings (Rivers, 2021). The primary intervention to promote self-testing developed in previous research is to enhance students' awareness of test-enhanced learning, such as providing scientific evidence of the testing effect or showing participants how much testing surpasses restudying according to their own learning experience (Bernacki et al., 2020). However, as demonstrated here, individuals with high TA may decline to test themselves for emotional reasons. Thus, we suggest that instructors identify students with

high TA and administer a validated TA intervention so that these students may actively use self-testing to enhance their learning. Additionally, previous studies have shown that frequent practice tests (e.g., class quizzes) can reduce students' TA (Agarwal et al., 2014; Khanna, 2015; Yang et al., 2021). Thus, it would be useful for instructors to have students engage in retrieval-based learning activities, especially ones in which success levels are high, by which they may become less anxious about testing and more willing to implement self-testing in their own study.

Several limitations of the present study should be noted. First, in Experiment 1, self-testing usage was only measured by a single survey item. Although the results are consistent with those in the other experiments, it is still necessary for future research to conduct large-scale investigations on this, not only to confirm the present findings, but also to explore potential moderating factors of the relation between trait TA and self-testing usage, such as self-efficacy and metacognitive awareness of test-enhanced learning.

Second, although differences in self-testing choices were observed between the high- and low-stake conditions in Experiments 3 and 4, it is unknown whether participants' state TA indeed differed between these two conditions because these experiments did not directly measure participants' state TA. Numerous studies have consistently confirmed that high-stake tests are more anxiety-provoking than low-stake ones, and the TA manipulation methods employed here have been repeatedly established as effective in previous studies (Hinze & Rapp, 2014; Wenzel & Reinhard, 2021). To address this limitation, we conducted Experiment 5 in which test stake was manipulated between-subjects, and state anxiety in both the high- and low-stake groups was measured. The results demonstrated a negative effect of state anxiety on self-testing choices. However, Experiment 5 also suffered from a limitation in that the TA measurement required participants to overtly report their concurrent anxiety feelings. Such an overt measurement of TA might reactively provoke awareness of TA and then change subsequent self-testing choices (Li et al., 2023; Shi et al., 2022; Zhao et al., 2023). Put differently, overt measurement of TA might induce confounds to the observed effect of TA on self-testing choices. Future research could measure TA in a more covert way, such as measuring participants' physiological arousal (e.g., skin conductance and heart rate) as a manipulation check (Roos et al., 2021).

The third limitation is that Experiments 2–5 measured self-testing usage by asking participants to make a binary choice between testing and restudying. Even though this is a widely used procedure to explore when and why learners choose to test themselves (e.g., Hui et al., 2021), such a binary decision is not fully representative of self-testing usage in applied settings. For instance, during daily self-regulated learning, students can make a choice among a variety of strategies, such as self-testing, rereading, note reviewing, concept mapping, and summarizing. Future research could profitably explore the effect of TA on self-testing usage by directly observing students' study behaviors rather than by simply asking them to make a forced choice between testing and restudying.

Concluding Remarks

High TA negatively affects self-testing usage. The detrimental effect of TA on self-testing works in both an inter- and intra-individual manner: Individuals with high trait TA are more likely to avoid self-testing compared with those with low trait TA; self-testing is dispreferred when individuals are preparing for a high-stake exam, even though the high-stake exam requires more frequent use of effective learning strategies (e.g., self-testing). Students may still decline to self-test despite appreciation of the benefits of testing, so the effectiveness of interventions aiming to enhance metacognitive beliefs may be limited, especially for individuals struggling with high TA. Self-testing avoidance is a potential factor explaining learning deficits caused by high TA. That is, due to avoidance of self-testing, individuals struggling with high TA may already be in a disadvantageous state when preparing for examinations before they actually undertake them.

Funding This research was supported by the Natural Science Foundation of China (32000742; 32171045; 32200841), the Research Program Funds of the Collaborative Innovation Center of Assessment toward Basic Education Quality at Beijing Normal University (2021–01-132-BZK01), the UK Economic and Social Research Council (ES/S014616/1), and the Fundamental Research Funds for the Central Universities (2022NTSS36).

References

Abu-Bader, S., & Jones, T. V. (2021). Statistical mediation analysis using the Sobel test and Hayes SPSS Process Macro. *International Journal of Quantitative and Qualitative Research Methods*, Retrieved March 25, 2024 from https://ssrn.com/abstract=3799204

Adesola, S. A., & Li, Y. (2018). The relationship between self-regulation, self-efficacy, test anxiety and motivation. *International Journal of Information and Education Technology, 8*, 759–763. https://doi.org/10.18178/ijiet.2018.8.10.1135

Agarwal, P. K., D'Antonio, L., Roediger, H. L., McDermott, K. B., & McDaniel, M. A. (2014). Classroom-based programs of retrieval practice reduce middle school and high school students' test anxiety. *Journal of Applied Research in Memory and Cognition, 3*, 131–139. https://doi.org/10.1016/j.jarmac.2014.07.002

Badali, S., Rawson, K. A., & Dunlosky, J. (2022). Do students effectively regulate their use of self-testing as a function of item difficulty? *Educational Psychology Review, 34*, 1651–1677. https://doi.org/10.1007/s10648-022-09665-6

Bartoszewski, B. L., & Gurung, R. A. R. (2015). Comparing the relationship of learning techniques and exam score. *Scholarship of Teaching and Learning in Psychology*, 1, 219–228. https://doi.org/10.1037/stl0000036

Bernacki, M. L., Vosicka, L., & Utz, J. C. (2020). Can a brief, digital skill training intervention help undergraduates "learn to learn" and improve their STEM achievement? *Journal of Educational Psychology, 112*, 765–781. https://doi.org/10.1037/edu0000405

Biwer, F., Egbrink, M. G., Aalten, P., & de Bruin, A. B. (2020). Fostering effective learning strategies in higher education—a mixed-methods study. *Journal of Applied Research in Memory and Cognition*, 9, 186–203. https://doi.org/10.1016/j.jarmac.2020.03.004

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417–444. https://doi.org/10.1146/annurev-psych-113011-143823

Blackwell, L. (2002). Psychological mediators of student achievement during the transition to junior high school: The role of implicit theories. *Columbia University*. Retrieved October 21, 2023 from https://www.proquest.com/docview/304792082?pq-origsite=gscholar&fromopenview=true

Bottiroli, S., Dunlosky, J., Guerini, K., Cavallini, E., & Hertzog, C. (2010). Does task affordance moderate age-related deficits in strategy production? *Aging, Neuropsychology, and Cognition, 17*, 591–602. https://doi.org/10.1080/13825585.2010.481356

Branscombe, N. R., & Wann, D. L. (1994). Collective self-esteem consequences of outgroup derogation when a valued social identity is on trial. *European Journal of Social Psychology, 24*, 641–657. https://doi.org/10.1002/ejsp.2420240603

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*, 1563–1569. https://doi.org/10.1037/a0017021

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21*, 279–283. https://doi.org/10.1177/0963721412452728

Cassady, J. C. (2004). The influence of cognitive test anxiety across the learning–testing cycle. *Learning and Instruction, 14*, 569–592. https://doi.org/10.1016/j.learninstruc.2004.09.002

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology, 27*, 270–295. https://doi.org/10.1006/ceps.2001.1094

Cipra, C., & Müller-Hilke, B. (2019). Testing anxiety in undergraduate medical students and its correlation with different learning approaches. *PLOS ONE*, *14*. https://doi.org/10.1371/journal.pone.0210130

Clark, D. A., & Svinicki, M. (2015). The effect of retrieval on post-task enjoyment of studying. *Educational Psychology Review, 27*, 51–67. https://doi.org/10.1007/s10648-014-9272-4

Costa, A., & Faria, L. (2018). Implicit theories of intelligence and academic achievement: A meta-analytic review. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00829

Culler, R. E., & Holahan, C. J. (1980). Test anxiety and academic performance: The effects of study-related behaviors. *Journal of Educational Psychology, 72*, 16–20. https://doi.org/10.1037/0022-0663.72.1.16

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques. *Psychological Science in the Public Interest, 14*, 4–58. https://doi.org/10.1177/1529100612453266

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review, 95*, 256–273. https://doi.org/10.1037/0033-295X.95.2.256

Endler, N. S., & Kocovski, N. L. (2001). State and trait anxiety revisited. *Journal of Anxiety Disorders, 15*, 231–245. https://doi.org/10.1016/S0887-6185(01)00060-3

Ergene, T. (2003). Effective interventions on test anxiety reduction. *School Psychology International, 24*, 313–328. https://doi.org/10.1177/01430343030243004

Ewell, S. N., Josefson, C. C., & Ballen, C. J. (2022). Why did students report lower test anxiety during the COVID-19 pandemic? *Journal of Microbiology & Biology Education*, *23*. https://doi.org/10.1128/jmbe.00282-21

Eysenck, M. W., & Calvo, M. G. (1992). Anxiety and performance: The processing efficiency theory. *Cognition & Emotion, 6*, 409–434. https://doi.org/10.1080/02699939208409696

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion, 7*, 336–353. https://doi.org/10.1037/1528-3542.7.2.336

Fairchild, A. J., & MacKinnon, D. P. (2008). A general model for testing mediation and moderation effects. *Prevention Science, 10*, 87–99. https://doi.org/10.1007/s11121-008-0109-6

Fan, T., Yang, C., & Luo, L. (2023). Improving the use of retrieval practice for both easy and difficult materials: The effect of an instructional intervention. In preparation for publication.

Geller, J., Toftness, A. R., Armstrong, P. I., Carpenter, S. K., Manz, C. L., Coffman, C. R., & Lamm, M. H. (2018). Study strategies and beliefs about learning as a function of academic achievement and achievement goals. *Memory, 26*, 683–690. https://doi.org/10.1080/09658211.2017.1397175

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review, 19*, 126–134. https://doi.org/10.3758/s13423-011-0181-y

Heitmann, S., Grund, A., Berthold, K., Fries, S., & Roelle, J. (2018). Testing is more desirable when it is adaptive and still desirable when compared to note-taking. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.02596

Hinze, S. R., & Rapp, D. N. (2014). Retrieval (sometimes) enhances learning: Performance pressure reduces the benefits of retrieval practice. *Applied Cognitive Psychology, 28*, 597–606. https://doi.org/10.1002/acp.3032

Hui, L., de Bruin, A. B. H., Donkers, J., & van Merriënboer, J. J. G. (2021). Does individual performance feedback increase the use of retrieval practice? *Educational Psychology Review, 33*, 1835–1857. https://doi.org/10.1007/s10648-021-09604-x

Hunsley, J. (1985). Test anxiety, academic performance, and cognitive appraisals. *Journal of Educational Psychology, 77*, 678–682. https://doi.org/10.1037/0022-0663.77.6.678

JASP Team. (2023). JASP (Version 0.17.1) [Computer software].

Johnson, M. L., & Sinatra, G. M. (2013). Use of task-value instructional inductions for facilitating engagement and conceptual change. *Contemporary Educational Psychology, 38*, 51–63. https://doi.org/10.1016/j.cedpsych.2012.09.003

Karatas, H., Alci, B., & Aydin, H. (2013). Correlation among high school senior students' test anxiety, academic performance and points of university entrance exam. *Educational Research and Reviews, 8*, 919–926. https://doi.org/10.5897/ERR2013.1462

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 331*, 772–775. https://doi.org/10.1126/science.1199327

Karpicke, J. D., & Roediger, H. L. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 704–719. https://doi.org/10.1037/0278-7393.33.4.704

Karpicke, J. D., Butler, A. C., & Roediger, H. L., III. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory, 17*, 471–479. https://doi.org/10.1080/09658210802647009

Karpicke, J. D. (2017). Retrieval-based learning: A decade of progress. In *Learning and Memory: A Comprehensive Reference*, 487–514. Elsevier. https://doi.org/10.1016/B978-0-12-809324-5.21055-9

Khanna, M. M. (2015). Ungraded pop quizzes: Test-enhanced learning without all the anxiety. *Teaching of Psychology, 42*, 174–178. https://doi.org/10.1177/0098628315573144

Kirk-Johnson, A., Galla, B. M., & Fraundorf, S. H. (2019). Perceiving effort as poor learning: The misinterpreted-effort hypothesis of how experienced effort and perceived learning relate to study strategy choice. *Cognitive Psychology, 115*, 101237. https://doi.org/10.1016/j.cogpsych.2019.101237

Kirkland, K., & Hollandsworth, J. G. (1980). Effective test taking: Skills-acquisition versus anxiety-reduction techniques. *Journal of Consulting and Clinical Psychology, 48*, 431–439. https://doi.org/10.1037/0022-006X.48.4.431

Koriat, A. (2008). Subjective confidence in one's answers: The consensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 945–959. https://doi.org/10.1037/0278-7393.34.4.945

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14*, 219–224. https://doi.org/10.3758/BF03194055

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory, 17*, 493–501. https://doi.org/10.1080/09658210902832915

Li, B., Zhao, W., Shi, A., Zhong, Y., Hu, X., Liu, M., Luo, L., & Yang, C. (2023). Does the reactivity effect of judgments of learning transfer to learning of new information? *Memory, 3*(1), 918–930. https://doi.org/10.1080/09658211.2023.2208792

Liebert, R. M., & Morris, L. W. (1967). Cognitive and emotional components of test anxiety: A distinction and some initial data. *Psychological Reports, 20*, 975–978. https://doi.org/10.2466/pr0.1967.20.3.975

Liu, S., Luo, L., & Yang, C. (2023). Test anxiety negatively predicts metacognitive awareness of test-enhanced learning and self-testing usage. In preparation for publication.

Ma, X., Li, T., Jia, R., & Wei, J. (2022). The forward testing effect in spatial route learning. *Acta Psychologica Sinica, 54*, 1433. https://doi.org/10.3724/sp.j.1041.2022.01433

McAndrew, M., Morrow, C. S., Atiyeh, L., & Pierre, G. C. (2016). Dental student study strategies: Are self-testing and scheduling related to academic performance? *Journal of Dental Education, 80*(5), 542–552. https://doi.org/10.1002/j.0022-0337.2016.80.5.tb06114.x

McCabe, J. A., & Lummis, S. N. (2018). Why and how do undergraduates study in groups? *Scholarship of Teaching and Learning in Psychology, 4*, 27–42. https://doi.org/10.1037/stl0000099

McDaniel, M. A., & Einstein, G. O. (2020). Training learning strategies to promote self-regulation and transfer: The knowledge, belief, commitment, and planning framework. *Perspectives on Psychological Science, 15*, 1363–1381. https://doi.org/10.1177/1745691620920723

Montoya, A. K. (2019). Moderation analysis in two-instance repeated measures designs: Probing methods and multiple moderator models. *Behavior Research Methods. 51*, 61–82. https://doi.org/10.3758/s13428-018-1088-6

Morris, L. W., & Fulmer, R. S. (1976). Test anxiety (worry and emotionality) changes during academic testing as a function of feedback and test importance. *Journal of Educational Psychology, 68*, 817–824. https://doi.org/10.1037/0022-0663.68.6.817

Myers, S. J., Davis, S. D., & Chan, J. C. K. (2021). Does expressive writing or an instructional intervention reduce the impacts of test anxiety in a college classroom? *Cognitive Research: Principles and Implications, 6*, 44. https://doi.org/10.1186/s41235-021-00309-x

Park, D., Gunderson, E. A., Tsukayama, E., Levine, S. C., & Beilock, S. L. (2016). Young children's motivational frameworks and math achievement: Relation to teacher-reported instructional practices, but not teacher theory of intelligence. *Journal of Educational Psychology, 108*, 300–313. https://doi.org/10.1037/edu0000064

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

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods, 162*, 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Putwain, D., & Daly, A. L. (2014). Test anxiety prevalence and gender differences in a sample of English secondary school students. *Educational Studies, 40*, 554–570. https://doi.org/10.1080/03055698.2014.953914

Putwain, D. W., & Symes, W. (2011). Achievement goals as mediators of the relationship between competence beliefs and test anxiety. *British Journal of Educational Psychology, 82*, 207–224. https://doi.org/10.1111/j.2044-8279.2011.02021.x

Putwain, D. W., & von der Embse, N. P. (2021). Cognitive–behavioral intervention for test anxiety in adolescent students: Do benefits extend to school-related wellbeing and clinical anxiety. *Anxiety, Stress, & Coping, 34*, 22–36. https://doi.org/10.1080/10615806.2020.1800656

Putwain, D. W., Connors, L., Symes, W., & Douglas-Osborn, E. (2012). Is academic buoyancy anything more than adaptive coping? *Anxiety, Stress, & Coping*, 25, 349–358. https://doi.org/10.1080/10615806.2011.582459

Putwain, D. W., Gallard, D., Beaumont, J., Loderer, K., & von der Embse, N. P. (2021). Does test anxiety predispose poor school-related wellbeing and enhanced risk of emotional disorders? *Cognitive Therapy and Research, 45*, 1150–1162. https://doi.org/10.1007/s10608-021-10211-x

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General*, 140, 283–302. https://doi.org/10.1037/a0023956

Rawson, K. A., & Dunlosky, J. (2012). When is practice testing most effective for improving the durability and efficiency of student learning? *Educational Psychology Review, 24*, 419–435. https://doi.org/10.1007/s10648-012-9203-1

Rea, S. D., Wang, L., Muenks, K., & Yan, V. X. (2022). Students can (mostly) recognize effective learning, so why do they not do it? *Journal of Intelligence, 10*, 127. https://doi.org/10.3390/jintelligence10040127

Rivers, M. L. (2021). Metacognition about practice testing: A review of learners' beliefs, monitoring, and control of test-enhanced learning. *Educational Psychology Review, 33*, 823–862. https://doi.org/10.1007/s10648-020-09578-2

Rivers, M. L., Dunlosky, J., & McLeod, M. (2022). What constrains people's ability to learn about the testing effect through task experience? *Memory, 30*, 1387–1404. https://doi.org/10.1080/09658211.2022.2120204

Robson, D. A., Johnstone, S. J., Putwain, D. W., & Howard, S. (2023). Test anxiety in primary school children: A 20-year systematic review and meta-analysis. *Journal of School Psychology, 98*, 39–60. https://doi.org/10.1016/j.jsp.2023.02.003

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 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*, 233–239. https://doi.org/10.1037/a0017678

Roick, J., & Ringeisen, T. (2017). Self-efficacy, test anxiety, and academic success: A longitudinal validation. *International Journal of Educational Research, 83*, 84–93. https://doi.org/10.1016/j.ijer.2016.12.006

Roos, A.-L., Goetz, T., Voracek, M., Krannich, M., Bieg, M., Jarrell, A., & Pekrun, R. (2021). Test anxiety and physiological arousal: A systematic review and meta-analysis. *Educational Psychology Review, 33*, 579–618. https://doi.org/10.1007/s10648-020-09543-z

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*, 1432–1463. https://doi.org/10.1037/a0037559

Sarason, I. G., & Stoops, R. (1978). Test anxiety and the passage of time. *Journal of Consulting and Clinical Psychology, 46*, 102–109. https://doi.org/10.1037/0022-006X.46.1.102

Segool, N. K., Carlson, J. S., Goforth, A. N., von der Embse, N., & Barterian, J. A. (2013). Heightened test anxiety among young children: Elementary school students' anxious responses to high-stakes testing. *Psychology in the Schools, 50*, 489–499. https://doi.org/10.1002/pits.21689

Shanks, D. R., Don, H. J., Boustani, S., & Yang, C. (2023). Test-enhanced learning. *Oxford Research Encyclopedia of Psychology*. https://doi.org/10.1093/acrefore/9780190236557.013.908

Shi, A., Xu, C., Zhao, W., Shanks, D. R., Hu, X., Luo, L., & Yang, C. (2022). Judgments of learning reactively facilitate visual memory by enhancing learning engagement. *Psychonomic Bulletin & Review, 30*, 676–687. https://doi.org/10.3758/s13423-022-02174-1

Soares, D., & Woods, K. (2020). An international systematic literature review of test anxiety interventions 2011–2018. *Pastoral Care in Education, 38*, 311–334. https://doi.org/10.1080/02643944.2020.1725909

Sotola, L. K., & Crede, M. (2021). Regarding class quizzes: A meta-analytic synthesis of studies on the relationship between frequent low-stakes testing and class performance. *Educational Psychology Review, 33*, 407–426. https://doi.org/10.1007/s10648-020-09563-9

Spielberger, C. D. (1980). *Test anxiety inventory: Preliminary professional manual*. Consulting Psychologists Press.

Spielberger, C. D., & Vagg, P. R. (1995). Test anxiety: A transactional process model. In C. D. Spielberger & P. R. Vagg (Eds.), *Test anxiety: Theory, assessment, and treatment* (pp. 3–14). Taylor & Francis.

Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R., & Jacobs, G. A. (1983). Manual for the state-trait anxiety inventory. Palo Alto, CA: Consulting Psychologists Press.

Steinmayr, R., Crede, J., McElvany, N., & Wirthwein, L. (2016). Subjective well-being, test anxiety, academic achievement: Testing for reciprocal effects. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01994

Stewart, D., Panus, P., Hagemeier, N., Thigpen, J., & Brooks, L. (2014). Pharmacy student self-testing as a predictor of examination performance. *American Journal of Pharmaceutical Education, 78*, 32. https://doi.org/10.5688/ajpe78232

Stowell, J. R., & Bennett, D. (2010). Effects of online testing on student exam performance and test anxiety. *Journal of Educational Computing Research, 42*, 161–171. https://doi.org/10.2190/EC.42.2.b

Sun, Y., Shi, A., Zhao, W., Yang, Y., Li, B., Hu, X., Shanks, D. R., Yang, C., & Luo, L. (2022). Long-lasting effects of an instructional intervention on interleaving preference in inductive learning and transfer. *Educational Psychology Review, 34*, 1679–1707. https://doi.org/10.1007/s10648-022-09666-5

Symes, W., & Putwain, D. W. (2020). The four WS of test anxiety. *Psychologica, 63*, 31–52. https://doi.org/10.14195/1647-8606_63-2_2

Szafranski, D. D., Barrera, T. L., & Norton, P. J. (2012). Test anxiety inventory: 30 years later. *Anxiety, Stress & Coping, 25*, 667–677. https://doi.org/10.1080/10615806.2012.663490

Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences, 110*, 6313–6317. https://doi.org/10.1073/pnas.1221764110

Theobald, M., Breitwieser, J., & Brod, G. (2022). Test anxiety does not predict exam performance when knowledge is controlled for: Strong evidence against the interference hypothesis of test anxiety. *Psychological Science, 33*, 2073–2083. https://doi.org/10.1177/09567976221119391

Thomas, C. L., Cassady, J. C., & Finch, W. H. (2018). Identifying severity standards on the cognitive test anxiety scale: Cut score determination using latent class and cluster analysis. *Journal of Psychoeducational Assessment, 36*, 492–508. https://doi.org/10.1177/0734282916686004

Tse, C.-S., Chan, M. H.-M., Tse, W.-S., & Wong, S. W.-H. (2019). Can the testing effect for general knowledge facts be influenced by distraction due to divided attention or experimentally induced anxious mood? *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00969

Tse, C.-S., & Pu, X. (2012). The effectiveness of test-enhanced learning depends on trait test anxiety and working-memory capacity. *Journal of Experimental Psychology: Applied, 18*, 253–264. https://doi.org/10.1037/a0029190

Toppino, T. C., LaVan, M. H., & Iaconelli, R. T. (2018). Metacognitive control in self-regulated learning: Conditions affecting the choice of restudying versus retrieval practice. *Memory & Cognition*, 46, 1164-1177. https://doi.org/10.3758/s13421-018-0828-2

Tullis, J. G., Finley, J. R., & Benjamin, A. S. (2013). Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. *Memory & Cognition, 41*, 429–442. https://doi.org/10.3758/s13421-012-0274-5

Tullis, J. G., Fiechter, J. L., & Benjamin, A. S. (2018). The efficacy of learners' testing choices. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*, 540–552. https://doi.org/10.1037/xlm0000473

Vaughn, K. E., & Kornell, N. (2019). How to activate students' natural desire to test themselves. *Cognitive Research: Principles and Implications, 4*, 35. https://doi.org/10.1186/s41235-019-0187-y

von der Embse, N., Jester, D., Roy, D., & Post, J. (2018). Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review. *Journal of Affective Disorders, 227*, 483–493. https://doi.org/10.1016/j.jad.2017.11.048

Weissgerber, S. C., & Rummer, R. (2023). More accurate than assumed: Learners' metacognitive beliefs about the effectiveness of retrieval practice. *Learning and Instruction, 83*, 101679. https://doi.org/10.1016/j.learninstruc.2022.101679

Wenzel, K., & Reinhard, M.-A. (2021). Does the end justify the means? Learning tests lead to more negative evaluations and to more stress experiences. *Learning and Motivation, 73*, 101706. https://doi.org/10.1016/j.lmot.2020.101706

Wissman, K. T., & Rawson, K. A. (2016). How do students implement collaborative testing in real-world contexts? *Memory, 24*, 223–239. https://doi.org/10.1080/09658211.2014.999792

Yan, V. X., Bjork, E. L., & Bjork, R. A. (2016). On the difficulty of mending metacognitive illusions: A priori theories, fluency effects, and misattributions of the interleaving benefit. *Journal of Experimental Psychology: General, 145*, 918–933. https://doi.org/10.1037/xge0000177

Yang, C., Potts, R., & Shanks, D. R. (2018). Enhancing learning and retrieval of new information: A review of the forward testing effect. *npj Science of Learning, 3*, 8. https://doi.org/10.1038/s41539-018-0024-y

Yang, C., Sun, B., Potts, R., Yu, R., Luo, L., & Shanks, D. R. (2020). Do working memory capacity and test anxiety modulate the beneficial effects of testing on new learning? *Journal of Experimental Psychology: Applied, 26*, 724–738. https://doi.org/10.1037/xap0000278

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin, 147*, 399–435. https://doi.org/10.1037/bul0000309

Yang, C., Li, J., Zhao, W., Luo, L., & Shanks, D. (2023). Do practice tests (quizzes) reduce or provoke test anxiety? A meta-analytic review. *Educational Psychology Review, 35*, 87. https://doi.org/10.1007/s10648-023-09801-wP

Yeager, D. S., Romero, C., Paunesku, D., Hulleman, C. S., Schneider, B., Hinojosa, C., Lee, H. Y., O'Brien, J., Flint, K., Roberts, A., Trott, J., Greene, D., Walton, G. M., & Dweck, C. S. (2016). Using design thinking to improve psychological interventions: The case of the growth mindset during the transition to high school. *Journal of Educational Psychology, 108*, 374–391. https://doi.org/10.1037/edu0000098

Yue, X. (1996). Test anxiety and self-efficacy: Levels and relationship among secondary school students in Hong Kong. *Psychologia: An International Journal of Psychology in the Orient*.

Zhao, W., Yin, Y., Hu, X., Shanks, D. R., Yang, C., & Luo, L. (2023). Memory for inter-item relations is reactively disrupted by metamemory judgments. *Metacognition and Learning, 18*, 549–566. https://doi.org/10.1007/s11409-023-09340-3

Weissgerber, S. C., Reinhard, M. A., & Schindler, S. (2018). Learning the hard way: Need for Cognition influences attitudes toward and self-reported use of desirable difficulties. *Educational Psychology*, 38, 176–202. https://doi.org/10.1080/01443410.2017.1387644

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Shaohang Liu[1,2] · Wenbo Zhao[3] · David R. Shanks[4] · Xiao Hu[2,5] · Liang Luo[1,2] · Chunliang Yang[1,2]

✉ Chunliang Yang
chunliang.yang@bnu.edu.cn

[1] Institute of Developmental Psychology, Faculty of Psychology, Beijing Normal University, Haidian District, 19 Xinjiekouwai Street, Beijing 100875, China

[2] Beijing Key Laboratory of Applied Experimental Psychology, National Demonstration Center for Experimental Psychology Education, Beijing Normal University, Beijing, China

[3] Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University, Beijing, China

[4] Division of Psychology and Language Sciences, University College London, London, UK

[5] Faculty of Psychology, Beijing Normal University, Beijing, China