5-2023

Meta-analysis of therapeutic interventions for the treatment of test anxiety.

Thomas Reece
Univirsity of Louisville

Recommended Citation
Reece, Thomas, "Meta-analysis of therapeutic interventions for the treatment of test anxiety." (2023). *Electronic Theses and Dissertations.* Paper 4072.
https://doi.org/10.18297/etd/4072

META-ANALYSIS OF THERAPEUTIC INTERVENTIONS FOR THE TREATMENT OF TEST ANXIETY

By

Thomas John Reece
B.A., Western Kentucky University, 2006
M.A., Western Kentucky University, 2009

A Dissertation
Submitted to the Faculty of the
College of Education and Human Development of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy
In Counseling & Personnel Services

Department of Counseling and Human Development
University of Louisville
Louisville, Kentucky

May 2023

META-ANALYSIS OF THERAPEUTIC INTERVENTIONS FOR THE TREATMENT OF TEST ANXIETY

By

Thomas John Reece
B.A., Western Kentucky University, 2006
M.A., Western Kentucky University, 2009

A Dissertation on

April 10, 2023

By the following Dissertation Committee:

____
Dissertation Director
Jeffrey C. Valentine

____
Myra Beth Bundy

____
Jason Immekus

____
Prathiba Natesan Batley

ACKNOWLEDGMENTS

I would like to thank my advisor Jeff Valentine. You have been remarkably patient and free with your time as I worked through this project. Thank you for your support. Thank you, as well to my committee members: Jason Immekus, thank you for the years of support during my time at UofL and for the continued encouragement to just get my dissertation done. Thank you to Prathiba Batley for agreeing to be on my committee, even though we have never actually had the chance to meet in person. Myra Bundy, thank you for agreeing to be my content expert for this dissertation. I always appreciate our opportunities to collaborate. *Herzlichen Dank*.

I also want to acknowledge Stephen Leach and Bo Yan for your assistance with the screening process for this project. You were generous with your time and your encouragement.

To John Bruni, thank you for being my first mentor in college and for showing me just how much fun statistics can be.

I also thank my wife Emily Reece, who beat me in the Ph.D. race, for all her support as I tried to catch up.

ABSTRACT

META-ANALYSIS OF THERAPEUTIC INTERVENTIONS FOR THE TREATMENT OF TEST ANXIETY

Thomas John Reece

April 10, 2023

This systematic review and meta-analysis was designed to be a practitioner-focused review of the current research into interventions for the treatment of test anxiety. As testing continues to be a large part of students' academic experiences and the stakes of that testing grow for students, teachers, and schools, there is a need for a synthesis of the literature to provide teachers and schools with some guidance on how best to help their students succeed. In this review, I describe the phenomenon of test anxiety and the current theoretical questions concerning the relationship between test anxiety and test performance. I also review prior syntheses of the test anxiety literature and describe why a new review is necessary.

After a robust literature search and double screening all reports found, eligible studies were double coded. Forty-two effect sizes, nested in 23 studies, were eligible for this review and reported sufficient information to be included in the meta-analysis. I conducted tests for publication bias and assessments of internal validity. The overall meta-analytic mean effect size was 0.22 standard deviations. Planned moderation tests explored the heterogeneity of the research base. Substantial, though not statistically significant, differences in effect sizes were noted for the type of intervention/therapeutic approach used, test subject, and academic level of the sample.

Overall, there are several limitations to the research base used in this meta-analysis. Underreporting of outcomes that were not statistically significant was common and studies often did not include basic information regarding the study sample. Approximately a third of the outcomes that met standards for inclusion in this review could not be included due to insufficient reporting of data needed to calculate an effect size. Additional problems related to the "light touch" approach of many interventions were also discussed, as was the seeming lack of theoretical foundations for some of the interventions. There is a need for additional research in this area to provide high-quality evidence of the effectiveness of test anxiety interventions on test performance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Anxiety refers to a general state of subjective feelings of tension, negative cognitions – as worry or apprehension – and physiological arousal that is interpreted in negative terms (American Psychiatric Association, 2022). Test anxiety is a specific type of anxiety related to the evaluative situation or process, or possible consequences of the evaluation (Sieber, O'Neil, & Tobias, 1977; Putwain, 2008; Gibson, 2014). There has been much research, and subsequent synthesis of that research, into predictors and correlates of test anxiety (von der Embse, 2018), but much less attention has been paid to the treatment of test anxiety. Prior syntheses of the test anxiety intervention literature have become dated (Ergene, 2003; Hembree, 1988) or limited (Huntley et al. 2019). With the increasing focus on testing for formative and summative purposes in education, and the subsequent potential for the misestimation of the ability level of students with test anxiety, there is a need for an updated analysis of interventions designed to treat test anxiety. To address this need, this systematic review and meta-analysis examines the current state of research in the field of test anxiety, specifically as it pertains to interventions designed to improve test performance. The ultimate aim of this review is to provide guidance to practitioners, teachers, and school leadership on how best to support students who experience test anxiety.

Estimates for the prevalence of severe or debilitating test anxiety among students range between 2% and 30% (Putwain & Daly, 2014) and it is probable that the number of students who experience less severe, but still problematic, levels of test anxiety is higher, up to 40% (von der Embse, 2013). The historical reliance on undergraduate college student populations for test anxiety research (Ergene, 2003) may result in an underestimate of the prevalence of severe test anxiety due to the effect of highly test anxious students self-selecting out of the college track (e.g., by choosing not to take required entrance exams). Also, due to the empirical and theoretical links between test anxiety and performance, test anxiety may represent a barrier to continuing education through lower than acceptable test scores (McDonald, 2001). Additionally, how test anxiety is operationally defined in surveys assessing its prevalence may affect this estimate. In an epidemiological study, Knappe, Beesdo-Baum, Fehm, Stein, Lieb, Wittchen, (2011) found 28% of participants expressed fear or anxiety of the testing situation (i.e., the performative element of testing, such as being monitored while engaging with the test), while 11% of the sample were specifically anxious about tests themselves (i.e., the summative purpose of tests, such as being evaluated based on one's performance on the test).

Reports of test-related worries generally increase with age (McDonald, 2001), with the average age for the onset of test anxiety being 14.7 years (Knappe, et al., 2011). This age coincides with entering high school in the U.S. and an increase in the stakes surrounding testing (e.g., high school exit exams, college entrance exams). The reports of test anxiety appear to be higher among female students, compared to their male counterparts across grade levels (Embse, 2018). Compared to white students, the prevalence of test anxiety among African American students is higher across all grade levels (Embse, 2018). Differences in test anxiety by race/ethnicity may be the result of systematic inequities in education found between racial groups, or an example of stereotype threat in action (Osborne, Tillman, & Holland, 2009).

The consequences for students who experience debilitating test anxiety can extend beyond the individual test. Test anxiety has been shown to be negatively correlated with self-esteem (*r* = -.42; von der Embse, 2018), and positively correlated with depression (*r* = .62; King, Mietz, Tinney, and Ollendick, 1995), other anxiety disorders (*r* = .20; Beidel, Turner, & Trager, 1994), poor class grades (*r* = -.18; Chapell et al., 2005), and a sense of helplessness regarding future assessments (*r* = .25; Cassady, 2004). Severe test anxiety may also result in test avoidance behaviors, where possible; avoiding challenging tasks in favor of simple or easy goals; or in self-handicapping, through procrastination or other forms of self-defeating behaviors, where the student builds a ready excuse for their poor performance, preserving their self-concept (Covington, 2000). Students who self-handicap may also use their test anxiety as an excuse, forming a self-fulfilling prophecy. Poor performance on a test reinforces a sense of the inevitability of future poor performance or failure, which can lead to self-defeating behaviors (Zeidner, 1998). The test anxious student may come to accept their inability to do well on tests. In a form of academic learned helplessness, such a student who has seen no improvement in performance perpetuates the trend of failure by no longer trying (Cassady, 2004). Avoidant behaviors could result in leaving formal education early (i.e., dropping out of school) or being reluctant to seek higher education for fear of standardized tests that often stand as gatekeepers (such the ACT/SAT or GRE) in the college admittance process (von der Embse, 2018).

Prior meta-analyses have found a relationship (correlations around -.20) between test anxiety and test performance (von der Embse, Jester, Roy, & Post, 2018; Hembree, 1988; Seipp, 1991). While small, this relationship could explain 4-5% of the variance in test performance and could mean the difference between passing and failing (or meeting some important score threshold) for many students. Given the proportion of scores clustered around the pass/fail threshold for the end of school exit exams in the UK (the General Certificate of Secondary Education, GCSE), Putwain (2008) estimated that a 4% variation in test performance attributable to test anxiety could be a determining factor for up to 20% of the students who pass or fail the exit exam in a given year. Test anxiety-related poor performance on college or program entry exams may also result in otherwise promising students having a disadvantage for admission to highly competitive programs, though a recent trend in dropping or making such exams optional has begun to remove this barrier (Jaschik, 2020). Once in postsecondary education, in certain highly rigorous or selective programs/majors, such as nursing, students who fail more than one course can be subject to dismissal from their program (Beggs, Shields, & Janiszewski Goodin, 2011).

Beyond the effect of test anxiety on the individual student, the suppression of test performance on aggregate could also affect the ratings of the school the student attends. In the US, the Every Student Succeeds Act requires public school accountability systems to use performance on state-mandated yearly assessments as part of performance evaluations of schools (Klein, 2015). A variety of negative consequences for schools, the school's administrators, and teachers, are associated with low performance on the state-mandated assessments, including a reduction in instructional autonomy (through the imposition of corrective action plans), termination of contracts for teachers or administrative staff, loss of performance pay or raises, etc. Overall, the consequences of test anxiety can be felt throughout the educational system and can affect the lives of those that rely on the results of testing.

Theories: What are the causes of test anxiety?

As described by Spielberger (1980) and, subsequently, by Spielberger and Vagg (1987), test anxiety includes components of state and trait anxiety. The anxiety is situation-specific and can be dependent on the characteristics of the test or testing environment, a characteristic of state anxiety. For example, tests that are perceived as high stakes or are described as difficult may elicit a greater anxiety response (Putwain, 2011, Von der Embse, 2018). Some conceptualizations of test anxiety emphasize the perception of threat present in the testing situation interacting with person-level traits and features of the testing situation (Zeidner, 1998). When participants are primed to think of the summative purpose of tests by describing a test's evaluative purpose, compared to when the test was portrayed as an activity or a learning scenario, participants reported greater levels of anxiety (Von der Embse, 2018). Spielberger and Vagg, as well as Zohar (1998), also suggest certain individuals may be predisposed to experiencing test anxiety or experiencing it to a greater degree (trait anxiety), which can in turn lead to reporting higher levels of anxiety during an exam (state anxiety). Zeidner (1998) argued that severe cases of test anxiety should be classified as a form of specific phobia, though to date test anxiety is not included as a recognized phobia in the DSM-5 (American Psychiatric Association, 2022). Others suggest the performative/evaluative elements of test anxiety most closely associates the condition with social anxiety (McDonald, 2001). LeBeau, Glenn, Liao, Wittchen, Beesdo-Baum, Ollendick and Craske (2010) echoed Zeidner's argument and added the practical concern that making test anxiety a diagnosable disorder would facilitate identification and treatment of the disorder, along with providing a basis for reasonable accommodations during testing.

What does text anxiety look like?

Symptoms of test anxiety can include cognitive symptoms such as rumination and worry related to either the testing situation or the possible results/consequences of the test, which may be distorted or perceived as being of greater significance than may objectively be the case, and/or the intrusion of task-irrelevant thoughts (Putwain, 2008). Along with the cognitive component of test anxiety, affective symptoms of anxiety, such as negative emotions, or physiological symptoms of anxiety and stress, such as increased perspiration and elevated heart rate, are commonly described (Putwain, 2008).

Does test anxiety affect performance?

Prior studies have found a negative correlation between test anxiety and test performance (Hembree, 1988; Seipp, 1991; Sommer & Arendasy, 2014). For example, Cassady and Johnson (2002) found that students who reported greater levels of anxiety during tests on a survey at the beginning of a course subsequently earned lower grades in their course and had lower self-reported SAT scores. There has, however, been some recent disagreement within the literature regarding the causal relationship between test anxiety and test performance (Sommer & Arendasy, 2014, 2015) and two competing interpretations of the anxiety-performance relationship have emerged. While both focus on characteristics of the situation, the direction of the causal arrows between ability, performance, and anxiety vary in important ways. In one model (the deficit hypothesis), there is no causal link between anxiety and performance and underlying ability level/preparation level causes both anxiety and performance, whereas the second model (the interference model) posits that anxiety directly affects or moderates performance (Zeidner, 1998).

The deficit hypothesis suggests students who report high levels of test anxiety are experiencing anxiety related to being unprepared or lacking necessary skills to successfully complete the assessment. Stated differently, the examinee rationally perceives themselves as lacking in some necessary capacity to do well on the test and this causes an increase in anxiety. While test anxious students tend to report studying for more hours a week than non-test anxious students, test anxious students also often report relying on less effective study strategies, such as repetition rather than elaboration, and surface-level processing of information (Cassady, 2009). Poor or ineffective study skills result in inefficient encoding, organizing, and storing of studied material, which in turn results in an inability to successfully retrieve the studied material during testing. The deficit model hypothesizes that the poor study habits demonstrated by test anxious students translates into poorer performance compared to non-test anxious students, regardless of how much longer the test anxious students spend studying (Cassady, 2009). There is also some evidence for the moderating effect of IQ on the relationship between test anxiety and test performance (McDonald, 2001), which would be expected based on the deficit hypothesis, but this effect is necessarily confounded because IQ is itself typically measured using a test. The crucial element of the deficit hypothesis is that the lack of ability or adequate preparation is the proximate cause of poor test performance and the anxiety experienced during the test session is simply an epiphenomenon.

The interference hypothesis conceptualizes the anxiety felt during testing itself as a moderator for test performance. In accordance with the inverted-U theory of arousal, anxiety can be beneficial to performance up to a certain optimal point but, beyond this point, more severe test anxiety can result in decreased performance (Yerkes & Dodson, 1908; Bodas & Ollendick, 2005). As anxiety becomes more elevated the test-taker may be prone to perceive the test or test-taking situation as threatening, diverting attention and cognitive resources needed to complete test items to worries or concerns related to their performance on the test. Preoccupation with irrelevant thoughts decreases the cognitive resources available for the task, taking the test, and information processing breaks down at the retrieval phase resulting in a biased estimate of the test-taker's true ability (Meijer, 2001). The anxiety may be related to a perceived lack in preparation/ability (real or imagined), an element of the testing situation, or it may be due to the stakes involved in the assessment. For example, one source of test anxiety and interference is perfectionism. A test taker who scores high on perfectionism may perceive a score less than 100% as "not good enough." This unreasonably high standard of performance may lead to a cycle of overly critical self-evaluations, anxiety and/or depression throughout the learning-testing cycle and may result in self-defeating behaviors. Alternatively, the test may be perceived as high stakes, such as an entrance exam or licensure exam, and the test taker may become preoccupied by the consequences of failure, leading to a similar cycle of catastrophizing, procrastination/avoidance of studying, and negative emotionality (Zeidner & Mathews, 2005). Regardless of the source of the anxiety, the interference hypothesis suggests it is the anxiety itself that hinders performance by interfering with information processing and recall.

From a psychometric perspective, test anxiety, as conceptualized by the interference model, may be considered a form of measurement bias (Haladyna &

Downing, 2004). Measurement bias refers to a situation in which individuals of the same ability level, or standing on a trait, score differently on a measurement instrument and the difference in their scores can be attributed to some factor unrelated to the construct the instrument is intended to measure. In this case, test anxiety is an unrelated third factor that may be a source of nonrandom error to a test. Taking this perspective, Sommer and Arendasy (2014) used a measurement invariance approach to examine test anxiety in cognitive tests. When test anxiety was used as the grouping factor, Sommer and Arendasy did not find evidence for a direct effect of test anxiety on test performance. The probability of a given score on the cognitive test was conditional solely on the underlying ability of the test taker and that probability was not affected by the test taker's level of test anxiety. In other words, they did not find evidence for a direct effect of test anxiety on test performance, which is in line with the predictions made by the deficit model. However, this study was conducted in a lab setting, so it is possible that the situation was not realistic enough, that the test was not difficult enough, or that the test was not perceived as sufficiently high stakes to provoke an anxiety response in high test anxiety test takers. Indeed, Hembree's (1998) meta-analysis found a stronger relationship between test anxiety and performance with difficult, in comparison to easy, tests, which seems to provide support for the deficit model. An easy test may be perceived as within the test taker's capabilities, even if they underestimate their actual ability level, and as such be less likely to prime the test taker's anxiety. However, the effect of perceived test difficulty on test anxiety may not provide evidence against the interference hypothesis because, though anxiety may still be present, the increased cognitive load from experiencing test anxiety may be less of an issue while taking an easy test. In a follow-up study to examine the possibility that the effects expected under the interference model may only be noticeable during high stakes testing situations, Sommer and Arendasy (2015) studied real-world applicants to medical school who were completing their admissions tests. Their study found that the addition of measures of test anxiety did not improve model fit during invariance testing, suggesting test performance was not conditional on level of test anxiety.

What are the methods of dealing with test anxiety?

The practical significance for the underlying cause of test anxiety is that each conceptualization of the role test anxiety plays in a testing situation dictates different treatments. The deficit model suggests the focus should be on improving test-related skills, and test anxiety is a symptom that will be reduced as skills improve, while the interference model sees anxiety as masking the test taker's true ability and, subsequently, implies anxiety-reduction techniques as a means to improving performance.

Hembree's (1988) meta-analysis suggests test performance has a stronger relationship with the cognitive components ($r = -.31$) of test anxiety than with the affective symptoms ($r = -.15$). Later studies also support this finding (e.g., Seipp, 1991, Sommer & Arendasy, 2014). This finding suggests therapeutic approaches that focus on treating cognitive symptoms (e.g., cognitive behavior therapy, mindfulness approaches) may be more effective than approaches that focus primarily on reducing affective symptoms (e.g., progressive muscle relaxation, systematic desensitization). Students who have a history of test anxiety accompanied by poor performance may, over time, develop a fixed mindset that they are "bad at tests" or school more broadly. Interventions targeting such students may need to first help the student combat their fixed self-concept before, or concurrently with, building their study/test-taking skills. For example, Mueller and Dweck (1998) found that students who were praised for their effort, as opposed to their performance, on a test (i.e., fostering a growth mindset concerning their intellectual/academic abilities) showed greater persistence after failure and a lower likelihood of challenge avoidance. Alternatively, the test may be interpreted as a threat in some way, possibly to one's self-esteem or status amongst peers, or to one's academic standing and all the follow-on consequences of meeting certain academic thresholds (Cassidy, 2004). In this case, an intervention may need to focus on reframing the test from a threat to a challenge or to foster a mastery orientation over performance (Davis, DiStefano, & Schutz, 2008). As can be seen, the treatment of test anxiety may not be a straightforward matter and a number of moderating factors may be at play.

Another wrinkle in treating test anxiety is that the stakes involved in testing can vary dramatically based on the age or educational level of the test taker. Von der Embse et al. (2018) found the correlation between test anxiety and grades varied by level (primary, intermediate, secondary, and post-secondary) with the smallest correlation being among secondary school students (r = -.16) and the largest correlation for post-secondary students (r = -.27). Similarly, Ergene (2003) examined the overall effect of test anxiety reduction interventions by education level (primary, middle, high, or college) and found large differences in effect sizes, with college and university samples having the largest effect size (d = .68) and secondary school having the smallest (d = .25). Examining the effect of test anxiety interventions, Hembree (1988) found some differences in effect sizes for the effect of treatment type on test anxiety that were moderated by academic level (college vs. precollege). Academic level is also a proxy for participant age and some therapeutic techniques may be more effective for some age groups than others due to differences in cognitive development. For example, there is debate as to whether children have the cognitive development necessary for cognitive behavioral therapy to be successful. If not, they would be better treated using a more behavior-focused approach (James, James, Cowdrey, Soler, & Choke, 2015).

What interventions have been developed to help students deal with test anxiety?

A wide variety of interventions have been proposed to reduce test anxiety including educational approaches such as tutoring (Faber, 2010) and test strategy training (Carter, 2005), and psychological approaches such as cognitive behavioral therapy (Yeo, 2016) and mindfulness (Jameson, 2014). More esoteric interventions, such as finger tapping (Zlomke, 2007), gum chewing (Ran et al., 2010), massage (Wettlaufer, 2017), isochronic tones (Pinnock, 2014), and acupuncture (Klausenitz, Hesse, Hacker, Hahnenkamp, & Usichenko, 2016) have also been proposed. With so many options of varying effectiveness, there is a need to synthesize the findings in the literature and provide clear guidance to practitioners who will be responsible for helping students overcome their test anxiety and improve their test performance.

What evidence do we have about the effectiveness of these interventions?

While there have been several meta-analyses examining correlates of test anxiety (Embse, et al., 2018, Hembree, 1988, Roos, et al, 2021, Seipp, 1991), to date, there have been only three prior meta-analyses that synthesized the effects of test anxiety interventions. Hembree (1988), as part of a larger ranging meta-analysis of test anxiety, examined the effect of various interventions for test anxiety. Importantly, Hembree reported the effect of the intervention on test anxiety and also on academic performance (test performance and GPA, separately). While at first glance this provides the opportunity to tease out the differential effect of various interventions on the two outcomes, the author appeared to not have this goal in mind when the analysis was published. Consequently, interventions were not necessarily grouped together in the same fashion. For example, systematic desensitization was an intervention for both outcomes; however, for the test anxiety outcome, separate effect sizes were calculated for middle/high school and postsecondary samples, while for the test performance outcome, only a single effect size is reported, representing samples from 6th grade through postsecondary. Insufficient information is provided in the text to calculate effect sizes that might be more comparable. To some extent differences in levels of aggregation are understandable, given the variety of studies that were included in the analysis and that there was no apparent intent to make comparisons between outcomes, but such comparisons would be valuable from a theoretical standpoint.

The next meta-analysis of test anxiety interventions synthesized the findings of test anxiety reduction intervention studies up to 1998 (Ergene, 2003). Ergene found a substantial degree of variability in effect sizes based on the intervention approach (i.e., cognitive-focused, skills-focused, etc.) and techniques (i.e., relaxation training, hypnosis, study skills training, etc.). When implemented alone, study skill-focused interventions showed a smaller average effect size estimate (d = 0.42), compared to cognitive therapy-only interventions which had a somewhat larger effect size estimate (d = 0.63), while interventions that combined cognitive therapy and skill-focused techniques resulted in the largest effect size (d = 1.22). The larger effect of the combined cognitive therapy and skills training interventions on the reduction of test anxiety symptoms suggests that the greatest improvements might be the result of multimodal interventions designed to treat irrational concerns about testing, while also bolstering study skills, thus reducing rational concerns about preparedness. The effect sizes reported by Ergene differed somewhat from those reported by Hembree (1988) for comparable intervention techniques. In general, effect sizes were in the same direction, but higher in Ergene's meta-analysis.

The most recent meta-analysis, by Huntley et al. (2019), is the only review which meets contemporary standards for conducting and reporting a meta-analysis. Effect sizes for test anxiety reduction interventions were similar to those reported by Ergene (2003), with the exception of study skills training, which was not statistically significant ($d = 0.02$), but there was only one study in this group. Study skills training was also not statistically significant in improving test performance ($d = 0.34$). There were only four studies within this category and the 95% confidence interval was large (ranging from -0.16 to +0.84), so this result may not be indicative of the true effect size for study skills training. The authors mention that the study skills focused intervention studies generally did not describe the content of the intervention in much detail, though they did note that the primary aim of these interventions was test anxiety reduction, rather than test improvement. If the reported effect size is representative of the true effect size, it is possible that study skills training was not effective because the intervention designers took an interference model, rather than a deficit model, approach in developing their training. It should be noted, however, that, just as was the case in the other meta-analyses, the interventions that combined study skills training with some other form of intervention (e.g., cognitive or behavioral therapy) had the largest effect size (anxiety reduction $d = 1.38$, performance improvement $d = 1.58$).

While Hembree (1988) and Huntley et al. (2019) synthesized the literature for test anxiety interventions that focused on test anxiety reduction and/or test performance improvement, Ergene (2003) focused solely on the effect of test anxiety interventions on the reduction of test anxiety symptoms and did not review studies or outcomes related to the effect of the intervention on actual test performance. While it was reasonable to assume interventions that have a positive effect on anxiety would have a positive effect on test performance, the question as to how much these interventions affect actual performance on tests was left unanswered. From a theory perspective, based on the previous discussion of the role that test anxiety may play in test performance and the uncertainty related to the strength of the relationship between the two constructs, it is important to examine both outcomes in a comprehensive synthesis of the literature. The current meta-analysis, however, will analyze specifically the test performance improvement literature. This is because, while understanding the relationship between anxiety reduction and performance improvement may be valuable to theoreticians, the current analysis is intended be practitioner-focused. By this I mean that the goal of this analysis is to provide evidence-based guidance to teachers, counselors, school/district administrators on how to support their test-anxious students in performing up to their potential.

Need for a state-of-the-art systematic review and meta-analysis

All three prior meta-analyses share some critiques in common related to the age of the literature in their reviews. The most recent studies included in Hembree (1988) were published in 1986, and 1998 for Ergene (2003), meaning the samples in the included studies come from a wholly separate generational cohort to students in grade, or even postsecondary, school today and the results may not be generalizable to the current generation of students. Additionally, Hembree's meta-analysis included studies from 1950 through 1986, with no comment made to the possible complications this may pose. Huntley et al.'s meta-analysis included studies from 1970 through 2017, with only six of the 44 included studies being published after 2000. Not only are multiple generations of students represented in these time frames, with no control for possible cohort effects or disaggregation by generation/decade, but the educational and therapeutic landscapes have also changed substantially over this time period. It is not clear a grade school sample from the 1950s is comparable to one from the 1980s or a sample from the 1970s being comparable to the 2000s.

In the United States, the emphasis on testing, such as through state-mandated standardized testing, has increased substantially since the late 1990s, particularly after the passage of the No Child Left Behind act in 2001 (Hart et al., 2015). Along with the increased use of testing in grade schools, the form tests take has also been changing, with a general shift away from paper-and-pencil tests and towards computerized testing. Additionally, with the continual changes in therapeutic techniques, such as the increasing popularity of mindfulness approaches in therapy in the past decade (Bellinger, DeCaro, & Ralston, 2015; Zenner, Herrnleben-Kurz, & Walach, 2014), the field has changed and the studies included in the prior meta-analyses have grown less relevant to current contexts.

Somewhat bridging the gap between Ergene's (2003) meta-analysis and the present time, Embse (2013) conducted a systematic review of test anxiety interventions published between 2000-2010. This review was limited in the scope of its search, however. Only peer-reviewed literature, thus ignoring grey literature, or "any document not issued by an entity with publishing as its primary purpose," (Young, Premji, & Englebert, 2021) was included in the search and only studies using K-12 samples were included in the review. As noted in Ergene, a large proportion of studies in test anxiety used college student samples, which might help explain why only 10 studies were included in Embse's review. While effect sizes were reported for each outcome measured by the included studies, the lack of a synthesis of the results makes interpreting the results difficult.

Huntley et al. (2019) resolves a number of the methodological issues noted earlier and provides an analysis that meets contemporary standards for reporting and rigor, however this most recent analysis can be criticized on other points. Huntley et al. restricted the scope of their analysis to only studies that used undergraduate student samples. This limits the usefulness of the study's results from a practitioner's viewpoint, as Huntley et al. themselves point out, because interventions that may be successfully employed in the postsecondary context may not be effective or feasible when working with K-12 students. Additionally, only randomized controlled trials that were published in peer-reviewed journals were included. While a defensible argument can be made for not including quasi-experimental studies, it does limit the pool of eligible studies and potentially excludes high quality studies that take place in more naturalistic settings, such as classrooms. The decision to exclude studies that were not published, however, is more problematic because of the 'file drawer' problem, in which studies which do not include a statistically significant outcome are less likely to be published and excluding null results inflates the observed effect size in a meta-analysis. The authors also allowed studies that used samples from outside the United States, as long as the study was written in English.

The inclusion of non-US samples introduces a complication because the context of testing can vary from culture to culture, both in the number of tests expected in a student's academic career and in the importance placed on such tests. For example, academic stress is a leading contributor to depression and suicide among South Korean students (Han & Lee, 2021). Additionally, attitudes toward the mental illness and psychotherapy can vary by culture (Kirmayer & Gomez-Carrillo, 2019) and may affect the effectiveness of interventions that are explicitly based on psychotherapy (e.g., cognitive therapy).

Stöber and Pekrun (2004) noted a decrease in the production of test anxiety research starting in the 1990s. This apparent decline in publication was attributed not to a decrease in research interest in the area, but rather to a focus on subject-specific forms of test anxiety, of which math anxiety is offered as an example of "hidden" test anxiety research. The existence of subject-specific test anxiety research poses a potential problem to any syntheses of test anxiety, as the question turns to whether these subcategories of test anxiety are in some way too distinct to be analyzed together. None of the currently published meta-analyses of test anxiety specifically include search terms that would allow the capture of subject-specific test anxiety research.

Math anxiety is perhaps the most commonly studied form of subject-specific test anxiety (Dowker, Sarkar, & Looi, 2016). Prior syntheses of test anxiety interventions have not included math anxiety terms in the database search. Hembree (1988, 1990) conducted separate meta-analyses for test anxiety and math anxiety, even while acknowledging that the two constructs are considered strongly related by many in the field (1990). Hembree's meta-analysis of math anxiety (1990) concluded with findings that ran parallel to those in his earlier study of general test anxiety (1988), though the possibility for the conceptual distinctiveness of the two was left open for further research. Recently, Caviola et al. (2021) conducted a meta-analysis of studies which examined the relationship of math anxiety or test anxiety on math performance. The meta-analytic correlation between math anxiety and math performance (*r* = -.30) and test anxiety and math performance (*r* = -.23) were similar. Importantly, in the subset of studies that measured both math anxiety and test anxiety, the meta-analytic correlation was considerable (*r* = .46).

Dowker et al. (2016) argued that math anxiety cannot be reduced to a subset of test anxiety, emphasizing the presence of anxiety related to math in non-testing contexts. Indeed, test anxiety and math anxiety can be defined in such a way as to be considered separate constructs, but in practice there is often enough overlap to justify including certain math anxiety studies in a synthesis of test anxiety interventions. As previously described, test anxiety has been defined as anxiety related to the of taking a test (Putwain, 2008) while the latter is anxiety associated with mathematics (Hembree, 1990). Though a distinction can be made that math anxiety may or may not occur while taking a test (e.g., a student with math anxiety may feel anxious during a math test and also in a non-evaluative setting, such as during a math lecture), the complication in the literature is that some studies of interventions for math anxiety are focused on decreasing anxiety experienced while taking a test or increasing performance on that test. What is being studied is anxiety while taking a test and, in such cases, whether the study's focus is referred to as "math anxiety" or "test anxiety" may simply be based on the researcher's preference or general line of research, rather than a substantive difference. Following

Gibson’s (2014) concept analysis of test anxiety, the defining attributes of test anxiety are a test of some sort and symptoms of anxiety. Test anxiety while taking a math test may be a subset of the broader construct of test anxiety, just as there may be other forms of content area-specific test anxiety. Including math anxiety in the search terms for this meta-analysis is also in keeping with the literature search strategy employed by the LeBeau et al. (2010) as part of the DSM-5 development work group responsible for recommending changes to the Specific Phobia diagnostic criteria, which included determining if test anxiety should be listed as a form of phobia.

While skill-building interventions that focus on developing content area proficiency may be specific to the content area, skill-building interventions focused on test-taking skills may be more general. It is likely that interventions intended to reduce anxiety, or increase performance, while taking a test will be similar regardless of the subject matter of that test. Similarly, the source of the anxiety (the testing situation itself or the content of the test) may not be relevant to the effectiveness of some interventions. For example, breathing exercises and progressive muscle relaxation work by countering the physiological manifestations of anxiety and the cause of the anxiety may not be as important with this intervention. Similarly, cognitive approaches to reducing anxiety (such as combating negative self-talk) may also be generalizable between the two concepts. The underlying cause of the anxiety may differ, and thus the specific negative beliefs being countered may differ (e.g., “I am bad at tests” versus, “I am bad at math”), but the manifestation of that anxiety, and its treatment, in the context of taking a test are likely to be similar. Whether the source of anxiety experienced while taking a test is the situation itself or the subject matter of that test is a matter that should be empirically examined. Additionally, after searching through the first 10 citations for math anxiety, three studies were identified that used math anxiety to refer to anxiety while taking a math test. At least two of these studies conflated math and test anxiety and used the term test anxiety in the text, but not in the abstract, and would have been missed in a search that did not include math anxiety in the search terms. Five studies examined math anxiety in ways that were not relevant to the current analysis (no test was involved) and two citations were generally irrelevant. Consulting a subject matter expert, they also expressed concerns about missing relevant studies because of inconsistencies in the way researchers conceptualize test and math anxiety (J.L. Adelson, February 9, 2018). For the purposes of this analysis, math terms were included in the search and test subject was coded, with the intention of conducting moderation analyses on math-related studies in comparison to non-math studies.

Overview

Test anxiety is experienced by many students at some point in their academic careers. For some, their experiences can be debilitating, for others, it may simply be unpleasant. On aggregate, the effect of test anxiety, both in terms of underestimating student ability and the avoidance of testing situations, can be consequential for students, their teachers and schools/school districts. Teachers and school administrators are in need of practical, research-based guidance for how to help students who experience test anxiety. Researchers are in need of an up to date and comprehensive review of the extant literature. While this review is designed to primarily to help the former, the latter will also benefit from seeing the current state of research and this review will, hopefully, serve as a guide for where research is still needed.

METHOD

Inclusion and Exclusion Criteria

To be included in the current analysis, studies had to examine the effect of an intervention focused on test anxiety whose outcome was improvement in test performance. Interventions could have been directed at individuals or groups and could have included a variety of forms (e.g., psychological, behavioral, or pharmacological treatments, study skills training, or some combination thereof). Studies of study skills or curriculum-based interventions had to include a high test anxiety sample (either by including results for a high test anxiety subsample or through sample selection criteria) to be included. Studies that focused solely on test anxiety reduction without measuring test performance were not included in the analysis. Studies could sample students of any age or grade level, starting at first grade, including students in postsecondary education. Studies which focused on test anxiety in teachers were not eligible.

Studies that took place in locations other than the United States were excluded because the focus of this meta-analysis is on interventions that may be generalizable to U.S. samples and each country's school system has its own testing regimes and emphasis on testing. Additionally, there is the potential of study results in non-U.S. samples being confounded by culture-specific practices. For example, some test anxiety interventions are based on psychotherapy approaches, and attitudes towards psychotherapy can vary widely across countries/cultures (Kirmayer & Gomez-Carrillo, 2019).

Only studies that included an intervention and at least one comparison group were included in the analysis. The comparison group had to be either business as usual, waitlist, or some other presumably nonefficacious intervention. The latter requirement means that single-group pre-post designs were not included, nor were studies that compared two interventions. Both quasi-experimental and randomized control trials were included. Post-test-only quasi-experimental designs were not included in the analysis because they could not control baseline differences between the intervention and comparison groups.

In line with What Works Clearinghouse (WWC) criteria for systematic reviews, which specify a document must be made publicly available within 20 years of the beginning of the review, all articles published between 2002 and 2022 were eligible for review. The rationale for this date restriction is that there is a concern that older studies may have taken place in contexts or with populations that are no longer as relevant to contemporary contexts (U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse, 2020a). This rationale is relevant for the current analysis: the field of psychotherapy has continued to change, with new therapeutic techniques being developed and older approaches falling out of favor. With the passage of the No Child Left Behind act in 2001, and the later passage of the Every Student Succeeds Act in 2015, there has been an increasing emphasis placed on state-mandated standardized testing in K-12 education, compared to before 2000. Testing, itself, has taken on new forms as more tests, either in the form of high-stakes tests like the ACT/SAT or simply tests taken as part of class, are administered digitally, rather than with paper/pencil. Finally, the U.S. student population (in grade school and in postsecondary) is growing more diverse and the samples used in older studies may not adequately reflect what may be considered the "typical" student in school today.

Information Sources

Search terms were developed with the assistance of a content expert. The search was designed to identify (a) empirical, quantitative studies of (b) the effects of interventions designed (c) to address test anxiety. The search sources are identified in Table 1 and the exact search terms are identified in Table 2. The test anxiety keywords were searched in the title, abstract, and subject terms. In addition, documents were searched without restriction for the research method indicators to identify studies that likely examined the effect of an intervention. A search limiter for date of publication was used to limit the scope of the search to studies that were published on or after 2002.

Table 1. *Platform and databases used in literature search*

Platform	Databases
EBSCO Academic Search Complete	o Academic Search Complete o Education Full Text (H.W. Wilson) o Educational Administration Abstracts o ERIC o MasterFILE Premier o OpenDissertations o Psychology and Behavioral Sciences Collection o APA PsycINFO o Social Sciences Abstracts (H.W. Wilson)
ProQuest	• Dissertations & Theses @ University of Louisville • EconLit • Ethnic NewsWatch • GenderWatch • ProQuest Dissertations & Theses Global • Sociological Abstracts

Table 2. *Search Strategy in EBSCO & ProQuest*

Search Number	Search Strategy	Concept Block
1	AB("test* anxiety" OR "math anxiety" OR "exam anxiety" OR "examination anxiety"	Test anxiety terms

	OR "mathematics anxiety" OR "test-anxious" OR "academic anxiety") OR TI("test* anxiety" OR "math anxiety" OR "exam anxiety" OR "examination anxiety" OR "mathematics anxiety" OR "test-anxious" OR "academic anxiety") OR SU("test* anxiety" OR "math anxiety" OR "exam anxiety" OR "examination anxiety" OR "mathematics anxiety" OR "test-anxious" OR "academic anxiety")	
2	FT(“control group*” OR random* OR “comparison group*” OR “matched group*” OR “treatment group*” OR experiment* OR evaluat* OR impact* OR effectiveness OR causal OR posttest OR post-test OR pretest OR pre-test OR QED OR RCT OR “propensity score” OR quasi-experimental OR efficacy OR “control condition*” OR “comparison condition*” OR “intervention group*” OR “intervention condition*” OR “no-intervention control” OR “wait-list” or “waitlist” or “waiting list” or waiting-list OR “intervention effect*” OR “intervention children” OR “control children” OR postintervention or post-intervention OR preintervention OR pre-intervention OR “treatment class*” OR “intervention class*”)	Research method indicators
3	1 AND 2	

In addition to searching databases, the reference lists of included articles were mined to identify any articles that may have been missed. The reference lists for Embse’s (2013) systematic review and Huntley et al.’s (2019) meta-analysis were used as a check on the sensitivity of the current search. For this sensitivity check, the current search was conducted without date limiters. All studies identified by Embse and Huntley et al. were found in the database search. A forward citation search using Google Scholar’s “cited by” function was also conducted on Embse’s (2013) systematic review and the meta-analyses by Ergene (2003), Hembree (1988), and Huntley et al. (2019) to identify any studies that were not found in the main document search.

A total of 5,825 citations were obtained from the database search and an additional 2,766 citations from the forward citation search of previous reviews and meta-analyses. Once citations were downloaded, they were loaded into citation management software. Duplicate citations were screened out using EndNote's 'Find Duplicates' function and the methods described by Bramer et al. (2016), resulting in 1,199 duplicate references being removed. During the course of document retrieval, an additional 19 references were identified as duplicates and removed.

Study Selection

Abstracts were screened by three screeners using Abstrakr (http://abstrackr.cebm.brown.edu). Abstrakr is an abstract screening tool that uses machine learning procedures to identify, and prioritize the screening of, abstracts that likely meet screening criteria based on previous abstracts that had been screened in (Wallace, Small, Brodley, Lau, & Trikalinos, 2012). Each abstract was reviewed independently by two screeners, me and one other screener. The two additional screeners both have experience with the whole process of systematic reviews and meta-analysis. Screeners were guided by questions addressing the inclusion criteria and were trained in screening procedures (see Table 3 for screening questions).

Table 3. *Study Screening Questions*

a) Does the study have an abstract written in English?

b) Does the study take place in the United States?

c) Does the publication examine the effect of an intervention?

d) Does the study implement an intervention whose key focus appears to be related to test anxiety?

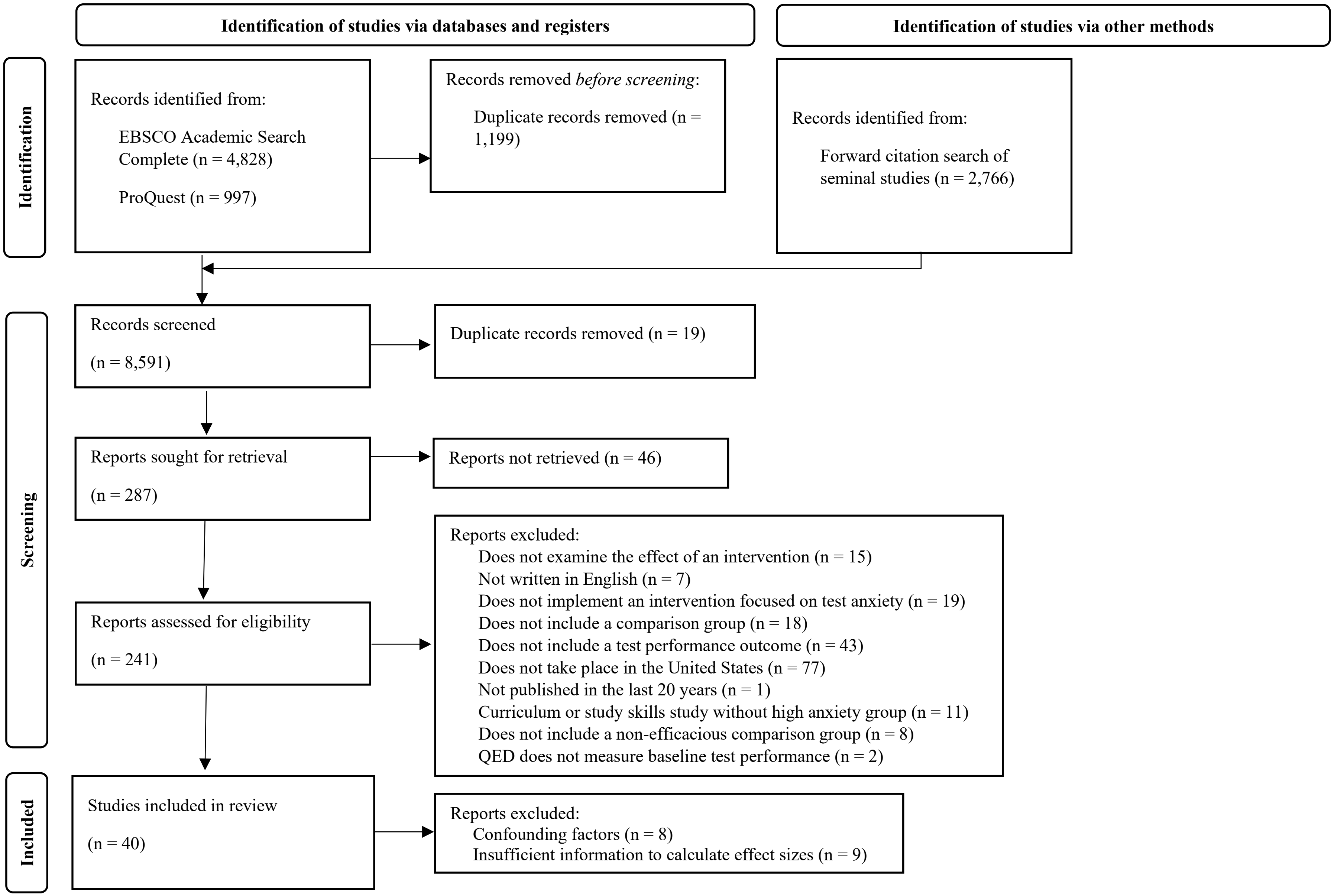
e) Does the study include a comparison group?

Screeners were instructed to screen out studies that clearly did not meet inclusion criteria. Studies were screened in if the abstract suggested that there was any form of comparison group. If the screener was unsure if an abstract met inclusion criteria (for example, because the abstract was vague, unclear, or omitted details required to make a decision) then the abstract was included for full text screening. After an initial pilot screening of 100 abstracts, the abstract screening guide was adjusted to provide greater clarity as to the inclusion criteria. After abstract screening was completed, screeners met to reconcile differences in abstract inclusion. Each screener described the rationale for their decision and if consensus could not be reached, the study was included for full text examination, in order to err on the side of inclusivity. Pre-reconciliation inter-rater agreement was 92%.

After abstract screening, all documents that appeared to meet inclusion criteria, or were not clearly ineligible for inclusion based on the abstract screening questions, were downloaded. Despite attempting to obtain each document using multiple methods (including through University-subscribed databases, Google Scholar, and requests through inter-library loan), a portion of the documents (46, 16%) were not obtainable. Of the documents that were not obtainable, 26 documents (56%) were likely to not be eligible based on the journal it was published in (i.e a regional journal such as the "Korean Journal of Youth Studies" or "Chinese Journal of Clinical Psychology") or the title or abstract were in English and a second language, suggesting the rest of the document may not be in English. The full text of the remaining 241 documents were then screened using the same inclusion criteria, with the additional requirement that the study must include a test performance outcome.

Figure 1 shows a flowchart of all identified studies for this analysis.

Figure 1. *Study Identification Flowchart*


Identification of studies via databases and registers
Identification of studies via other methods
Identification
Records identified from:
EBSCO Academic Search Complete (n = 4,828)
ProQuest (n = 997)
Records removed *before screening*:
Duplicate records removed (n = 1,199)
Records identified from:
Forward citation search of seminal studies (n = 2,766)
Screening
Records screened
(n = 8,591)
Duplicate records removed (n = 19)
Reports sought for retrieval
(n = 287)
Reports not retrieved (n = 46)
Reports assessed for eligibility
(n = 241)
Reports excluded:
Does not examine the effect of an intervention (n = 15)
Not written in English (n = 7)
Does not implement an intervention focused on test anxiety (n = 19)
Does not include a comparison group (n = 18)
Does not include a test performance outcome (n = 43)
Does not take place in the United States (n = 77)
Not published in the last 20 years (n = 1)
Curriculum or study skills study without high anxiety group (n = 11)
Does not include a non-efficacious comparison group (n = 8)
QED does not measure baseline test performance (n = 2)
Included
Studies included in review
(n = 40)
Reports excluded:
Confounding factors (n = 8)
Insufficient information to calculate effect sizes (n = 9)

Data Collection and Study Coding

Two coders, working independently, completed a standard coding protocol for each study. The study coding guide was developed by the author and tested by both study coders during a pilot round of coding several studies. The coders then met to discuss ways of improving the coding guide and identified other study characteristics that should be captured. Data collected during study coding included data on participant characteristics (gender, age, ethnicity, academic level), study data (year and type of publication, study setting, sample size, attrition and/or baseline equivalence data, measurement instruments used), intervention characteristics (technique, length of intervention, comparison type), and statistics relevant to computing effect sizes. A version of the coding guide, along with all data and code used in this synthesis, is available in the GitHub repository linked in Appendix A.

During the study coding process, coders met regularly to discuss coding questions and to ensure there was consistency in study coding. All discrepancies between coders were settled in discussion. Additional study characteristics were added to the code sheet as themes emerged to allow for further moderation analyses. When a study included the same outcome measured at multiple time points, each time point was coded separately so that the effect the intervention at different time intervals postintervention could be examined. Pre-reconciliation inter-rater agreement was 99% across all coded fields.

A single article or dissertation may report multiple studies. For the purpose of this analysis, multiple studies within a single manuscript were considered separate studies if the samples were non-overlapping. For example, a document that included results separately for several different school districts, or multiple samples that were randomly assigned separately into non-overlapping groups, may be considered as containing several studies, but a subgroup analysis of a single sample would not. When a study reported subgroup analyses in addition to a whole group analysis, the results for the full sample was preferred. Subgroup analyses were coded as supplemental outcomes for possible moderator analyses. If only subgroup analyses were reported and sufficient information was provided to calculate an overall effect size, the overall effect size was calculated by the coders. A single study may be reported in multiple publications (e.g., a dissertation and a subsequent journal article, or a single sample followed over time in multiple articles). In these cases, the publications were treated as a single study for the purposes of study coding and analysis and the more detailed report was preferred for coding purposes.

Methods for Assessing Risk to Internal Validity

Study Quality Assessment

Attrition and Statistical Control. Study quality was assessed by coding elements of the study's research design. For randomized controlled trials (RCTs), overall and differential attrition were assessed. Attrition is the greatest threat to the internal validity of RCTs (Shadish, Hu, Glaser, Kownacki, & Wong, 1998), as sample loss can compromise the expected statistical similarity of groups in the study and, thus, potentially bias the results of the study (U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse, 2014a). Differential attrition can be particularly problematic because such attrition may be related to some aspect of the study that is unique to one group. To determine if attrition was significant in an RCT, the attrition standards detailed in the WWC *Standards Handbook* version 4.1 (U.S. Department of

Education, Institute of Education Sciences, What Works Clearinghouse, 2020a) were used. The WWC standards establish two thresholds (one based on more cautious assumptions, one based on more optimistic assumptions about attrition) for assessing attrition, based on the assumptions made about the potential bias that might arise due to attrition. I used the cautious threshold for this meta-analysis because I believe that, when dealing with treatment for anxiety, it is very likely that participant loss will be related to the intervention itself. In addition to attrition in RCTs, I coded for the presence of participants who were added to the analytic sample after random assignment or participants that were reassigned to a different group after random assignment, as these represent a potential compromising of the random assignment. RCTs with high attrition or whose randomization was compromised were treated as if they were QEDs. I identified 1 RCT study (1 outcome) with high attrition and this study was treated as a QED. Eleven studies (21 outcomes) did not provide sufficient information to calculation attrition and were included as supplemental studies to the main analysis. There were no RCT studies with compromised randomization.

To be included in the analysis, QED studies had to control for pre-intervention test performance. Controls could include subtracting the pretest effect size from the posttest effect size (either reported by the study or performed during study coding) or include the pretest scores in the analytic model. If adjustments for baseline differences were needed, but not done by the study authors, and sufficient information was provided in the document, the appropriate adjustments were made during study coding (e.g., subtracting the baseline effect size from the post-test effect size). Whether the authors of a QED study controlled for pre-intervention test anxiety was coded and I planned to conduct a moderator analysis to determine if the effect sizes between studies that controlled for baseline differences in test anxiety level were substantially different. Of the 3 QED studies included in the analysis, none of them controlled for baseline test anxiety.

Confounds. Studies that included confounding factors in the analysis were not included in the meta-analysis as their results are not sufficiently credible. Examples of confounding factors include: a) the intervention and/or comparison group consisted of only one class/teacher/school each (in this case, there is no way to determine if the observed effect is the result the intervention or a characteristic of the grouping factor) or b) if there was a variable unrelated to the intervention that is only present in one group (in this case, the effect of the intervention cannot be disentangled from the unrelated variable). I excluded 7 studies due to confounding factors. Five studies had intervention and/or comparison groups that consisted of one teacher or school and the groups were non-overlapping (Donato, 2010; Dreisbach, 2017; Leap, 2013; Lobman, 2014; Wisinger, 2010). One study was excluded because group assignment was based on the participants' belief in the malleability of intelligence (growth mindset; a potentially endogenous factor; Wieland, 2011). The last study to be excluded used final course grades as a substitute for test performance where test scores were not available (Driscoll, Holt, & Hunter, 2005).

Measurement of Test Anxiety

Validity

A valid measure is one that clearly measures the construct it was designed to measure. Standardized assessments and tests of test anxiety that have published validation studies were considered reliable and valid if they were used unmodified and in a manner consistent with their intended use. There are a handful of widely used measures of test anxiety for example, the Spielberger Test Anxiety Inventory (Spielberger, 1980), and the Math Anxiety Rating Scale (Suinn, Edie, Nicoletti, & Sinelli, 1972) whose psychometric properties have been well researched. Well-established measures of test anxiety were always included. Other measures that purport to measure test anxiety were judged by the reviewer on their face validity based on the full item set and were excluded if (a) they did not appear to be face valid or (b) they did not provide enough information to make a face validity judgement. If a study used an established test anxiety measure which was modified from the original, it was treated as a new measure and the coder made a judgement as to whether the stated changes threatened the validity of the study results. The coding guide captured information about the instruments used to measure test anxiety and test performance in the study and whether they were modified.

Reliability

A reliable measure is one which is internally consistent. Reliability information, such as coefficient alpha, for the study sample was collected and reported, but poor reliability was not an exclusion criterion for this analysis. The primary effect of low reliability is to reduce the observed treatment effect in the study. Removing studies that report low reliability may have the effect of inflating the treatment effect calculated in this meta-analysis, whereas retaining these studies provides a more conservative estimate. Coefficient alpha, or equivalent, was recorded if the study reported study-specific reliability.

Summary Measures

The standardized mean difference, adjusted for small sample size bias (Hedges' *g*; Hedges, 1981), was used as the measure of effect size in the meta-analysis:

$$g = \left(\frac{\bar{Y}_1 - \bar{Y}_2}{s_p}\right) * (1 - \frac{3}{4n - 9})$$

Where *n* is the total sample size for the analytic sample, $\bar{Y}$ is the group mean at post-test, and s_p is the pooled standard deviation at post-test. When statistical adjustment was required, effect sizes were computed using adjusted means and unadjusted standard deviations. I calculated effect sizes when sufficient information was reported in the study. Where group means and standard deviations were not available in the study, but other data were reported, the formulas found in Shadish, Robinson, and Lu (1997) were used to calculate an effect size with the available information. Study-reported effect sizes were only used when insufficient information was provided and the reported effect size appeared to be comparable to Hedges' *g*. When study-reported effect sizes were used but the analytic sample size was not provided by condition, sample sizes for each condition were estimated by evenly dividing the overall sample size in order to calculate the standard error for the effect size.

When only gain score standard deviations are reported, they were adjusted by the baseline-outcome correlation using the procedures outlined by the WWC *Standards Handbook* (U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse, 2020a). If the baseline-outcome correlation was not reported, a correlation of .80 was used, which provides a conservative estimate of the underlying relationship between time points. Standard deviations were estimated in this manner for one study (Bishop, 2007).

Methods of Synthesis

All analyses were run in R (version 4.1.0; R Core Team, 2017), using the "metafor" package (version 3.4-0; Viechtbauer, 2010). When synthesizing the results of multiple studies in a meta-analysis, a conceptual concern is whether to treat each individual study functionally as replication studies estimating the same population parameter (i.e., a single true effect size) or as sampling from a population of effect sizes. In the former case, fixed effects meta-analysis, which assumes that all studies are sampling from the same population and differences between observed effect sizes are due to sampling error, would be appropriate. If, however, the fixed effects assumption is not tenable (due to anticipated between-study variability in samples and procedures), then random effects meta-analysis is preferred. Random-effects meta-analysis assumes that study effect sizes come from a larger population of effect sizes, which are distributed about a mean effect size and variation in study effect sizes are the result of sampling error at the subject level as well as at the study level (Borenstein, Hedges, Higgins, & Rothstein, 2009). Given the great deal of variability in procedures and samples between studies in this analysis, one cannot assume these studies represent direct replications of each other and it is reasonable to make the random-effects model assumption. For this reason, I used random effects meta-analysis in the current synthesis and meta-analytic weights were based on the inverse of the random effects variance of the effect size.

When multiple comparisons were included in a study, or where multiple interventions were compared to a control, I planned on addressing non-independent effects were addressed by using robust variance estimation using the "robumeta" package (version 2.0; Fisher, Tipton, & Zhipeng, 2017) but ultimately used two different approaches as described below. Where a study reported a delayed and immediate post-test condition, the immediate condition was considered the primary outcome to ensure consistency of outcomes across studies. Delayed conditions were considered secondary outcomes and were coded for potential moderator analysis.

I conducted a moderation test using research design (RCT, QED) as the moderator to determine if RCT and QED studies differed in the size of their effects and whether these studies should be combined in the meta-analysis. There was not a substantive difference between studies based on research design, so RCT and QED studies were combined in the analysis. Moderator analyses were conducted to determine if the average effect size varied by academic level (elementary, secondary, postsecondary), therapeutic approach (behavioral, cognitive, study skills training, combined), and test subject (math, language arts, other).

Publication Bias and Selective Reporting

A meta-analysis is only accurate when the studies included in the analysis represent the total population of studies that meet the inclusion criteria, either in its entirety or as a random subset. Publication bias refers to when there exist studies which would meet inclusion criteria that remain to be uncovered and these studies systematically differ from those included in the meta-analysis. In this analysis, publication bias was addressed by using a wide range of approaches, including the tests provided by the “metafor” package. The specific publication bias assessments used included moderation tests where publication status was the moderator, the production of funnel plots, and statistical tests of asymmetry. Because tests of publication bias assume that effect sizes are independent, where there are multiple effect sizes per study, an aggregated effect size (and an adjusted standard error that reflects the additional power gained from having multiple measures of the same construct) was calculated using the ‘aggregate’ function provided in the “metafor” package. For the purposes of examining publication bias, only the immediate posttest outcomes (where outcomes at multiple time points were reported) were included.

Publication status was used as a moderator in a moderator analysis to determine if there was a significant difference in average effect size by publication status. A significant difference in the size of the effect found in published and unpublished studies suggests there may be systematic differences between the two groups of studies and is therefore another way to explore potential publication bias. Additionally, funnel plots, in which the effect size for each study is plotted against the standard error, were produced. In general, effect sizes with lower standard error will tend to cluster together near the mean effect size and, as the standard error increases, the spread of effect sizes will become more pronounced. In the absence of publication bias, the effect sizes in the analysis should be symmetrically distributed about the mean effect size. A lack of symmetry, particularly as error increases, would indicate the potential for publication bias. Visual examination of the funnel plot was supplemented with Egger’s regression test, in which the effect size is regressed on the standard error, (Egger, Smith, Schneider, & Minder, 1997) and Begg and Mazumdar’s rank correlation test, in which the correlation between the effect size and variance is calculated, (Begg & Mazumdar, 1994) to provide formal tests of funnel plot asymmetry. For both tests, a statistically significant result indicates the possibility of publication bias. Additionally, the trim and fill procedure was used (Duval and Tweedie, 2000). In this procedure, the studies contributing the most to asymmetry in the funnel plot were iteratively trimmed, until the plot was symmetric. A new effect size was then calculated, and the trimmed studies were added back to the plot, along with mirror images of those studies. A comparison was then made between the original mean effect size and the trim and fill effect size.

RESULTS

Description of Included Studies

After screening and coding 8,591 citations, 40 reports met eligibility requirements for this analysis. Of these 40 reports, 8 were excluded due to confounding factors. An additional 9 were excluded because they did not include sufficient information to calculate effect sizes. Table 4 shows sample characteristics for all studies included in the analysis. In this table, along with tables 5-7, the table is divided into three sets of studies (as indicated in the “Analysis Type” column). The “Main” studies shown at the top of the table represent the studies included for the primary analyses. These are the 12 studies that met all eligibility and study quality requirements and included the required data to calculate usable effect sizes. The studies labeled “Supp” include an additional 11 RCT studies that did not include sufficient data to calculate attrition, but otherwise met eligibility and quality requirements. The studies labeled “No ES” are the studies that did not include sufficient information to calculate effect sizes.

As can be seen in Table 4, few studies provided all the sample background information coded in the study. Most notably, whether the sample was drawn from an urban, suburban, or rural setting was rarely reported. Mean/median ages were also often not reported; instead either an age range or, for K-12 samples, student grade was often provided. The majority of studies were conducted in a classroom setting using classroom exams or some form of standardized test. Most samples were small (ranging from 10-791 with an interquartile range of 24-81)

Table 4. *Sample Characteristics*

Study Author(s) and Year	Study Inclusion	Setting	Urbanicity	% Male	Age	Academic Level	Race/Ethnicity	Type of test	Analytic sample overall N	Subject of test
Bishop, 2007[a]	Main	Classroom	Rural	40%	10th grade	HS	NS	State test	30	ELA, Math
Dolton, 2016	Main	Classroom	NS	77%	9.18	ES	"Ethnic groups that were represented included Caucasian and Hispanic" (p. 41)	Standardized test	22	ELA, Math
Evans et al., 2010[a]	Main	Classroom	NS	NS	NS	PS	NS	Standardized test	42	Nursing
Harrison, 2016	Main	Classroom	Urban	NS	14-18	HS	NS	Standardized test	18; 22	ELA, Science, Composite
Haynes, 2003	Main	Classroom	NS	NS	NS	PS	NS	Classroom exam	160	Math
Hines, 2011	Main	Classroom	NS	NS	16.56	HS	76% African-American/Non-Hispanic, 18% Caucasian/Non-Hispanic, 3% Hispanic, 1% Asian, 2% Multiracial	State test	93	Math
Huang and Mayer, 2016	Main	Lab	NS	31%	18.55	PS	NS	Lab-based test	54	Math
Husni, 2006	Main	Classroom	Rural	NS	18-19	PS	100% African American	Institutional exam	60	Math
Jamieson et al., 2016	Main	Classroom	Urban	31%	29.4	PS	69% Black/African American and 31% Caucasians (p. 3)	Classroom exam	81	Math

Table 4. *Sample Characteristics*

Study Author(s) and Year	Study Inclusion	Setting	Urbanicity	% Male	Age	Academic Level	Race/Ethnicity	Type of test	Analytic sample overall N	Subject of test
Namwamba, 2013	Main	Lab	NS	77%	18-30	PS	NS	Lab-based test	12	Math
Nelson and Knight, 2010	Main	Classroom	NS	35%	NS	PS	NS	Classroom exam	118	Psychology
Spielberger, 2015	Main	Classroom	NS	35%	20.86	PS	63% Caucasian, 12% Asian, 8% African American, 6% Hispanic, 1% Native American, 5% Other, and 6% two or more races	Classroom exam	106	Psychology
Brady et al., 2017	Supp	Classroom	NS	42%	NS	PS	21% multiple racial-ethnic identities, 64% White, 27% Asian, 11% Black, 14% Hispanic or Latino/a, 7% Native American or Pacific Islander, and 1% Other	Classroom exam	194; 237	Psychology
Falcon, 2017	Supp	Classroom	Urban	53%, 58%	12, 13	MS	"predominantly Hispanic-Latino (90%)", (p. 51)	State test	42; 53	ELA
Goldenberg et al., 2013	Supp	Classroom	Urban	34%	20.34	PS	51% White or Caucasian, 27% Asian, 14% Other, 6% Black or African American, and 2% Native Hawaiian or Other Pacific Islander.	Classroom exam	176	Psychology
Im, 2013	Supp	Lab	NS	58%	24.07	HS	81% African-American, 10% White/Caucasian, 6%	Lab-based test	41; 42	Math

Table 4. *Sample Characteristics*

Study Author(s) and Year	Study Inclusion	Setting	Urbanicity	% Male	Age	Academic Level	Race/Ethnicity	Type of test	Analytic sample overall N	Subject of test
							Asian/Pacific Islander, 4% Hispanic/Latino			
Insalaca, 2007	Supp	Classroom	NS	NS	14-18	HS	NS	Classroom exam	80	Math
Kim et al., 2017[a]	Supp	Classroom	NS	52%	15.91	HS	NS	Assessments developed for the curricular content	32	Math
Park et al., 2014[a]	Supp	Lab	NS	NS	NS	PS	NS	Lab-based test	44; 42	Math
Perez, 2005	Supp	Classroom	NS	38%	18-21	PS	100% Hispanic/Latino	State test	123	Math
Shen, 2009[a]	Supp	Lab	NS	NS	16+	HS	70% African American	Lab-based test	50; 55; 56	Math
Shobe et al., 2005	Supp	Lab	NS	10%	21.7	PS	NS	Lab-based test	10	Math
Thompson et al., 2016	Supp	Classroom	NS	55%	10.65	ES	7% African American, 45% Asian American, 39% Latino, 6% White, 3% Other	Standardized test	791; 709	Math, ELA
Blank-Spadoni, 2013	No ES	Classroom	NS	43%	NS	PS	50% White, 8% African-American, 15% Hispanic/Latino, 9% Asian,	Classroom exam	158	Mixed

Table 4. *Sample Characteristics*

Study Author(s) and Year	Study Inclusion	Setting	Urbanicity	% Male	Age	Academic Level	Race/Ethnicity	Type of test	Analytic sample overall N	Subject of test
							3% Hawaiian/Pacific Islander, 15% Other/2 or More Races			
Edwards, 2012	No ES	Classroom	NS	68%, 51%	NS	PS	31% White, 35% Black, 20% Hispanic, 2 Islander, 12 Asian	Classroom exam	49	Math
Edwards, 2012	No ES	Classroom	NS	68%, 51%	NS	PS	74% White, 9% Black, 15% Hispanic, 1 Islander	Classroom exam	30	Math
Ganzenmuller, 2007	No ES	Classroom	Urban	48%	NS	Undergraduate, Graduate	NS	Standardized test	24	ELA
Harris et al., 2019	No ES	Classroom	Urban	35%	NS	PS	NS	Classroom exam	779	Biology
Harrison, 2016	No ES	Classroom	Urban	NS	14-18	HS	NS	Standardized test	NA	ELA, Science, Composite
Henslee & Klein, 2017[a]	No ES	Lab	NS	NS	20	PS	NS	Lab-based test	58	Math
Insalaca, 2007	No ES	Classroom	NS	NS	14-18	HS	NS	Classroom exam	80	Math
Jacobs, 2021	No ES	Classroom	NS	54%	22.6	PS	48% Caucasian, 16% African American, 9% Asian, 21% Latino, and 6% Other	Standardized test	228; 229	Math

Table 4. *Sample Characteristics*

Study Author(s) and Year	Study Inclusion	Setting	Urbanicity	% Male	Age	Academic Level	Race/Ethnicity	Type of test	Analytic sample overall N	Subject of test
Miller et al., 2006[a]	No ES	Classroom	Rural	44%	5th grade	ES	"Primarily Caucasian" (p. 5), 25% African American and 3% Hispanic	State test	36	Composite, ELA, Math, Science, Social Studies
Miller et al., 2007[a]	No ES	Classroom	Rural	34%	6th grade	MS	NS	State test	61	Composite
Sefton, 2014	No ES	Classroom	NS	52%	21	PS	49.4% White, 43.8% African American, 4.5% Asian, 2.2% Hispanic	Classroom exam	64	Math

ES = Elementary School, MS = Middle School, HS = High School, PS = Postsecondary

NS = Not Specified

ELA = English Language Arts

Supp = Supplemental Studies

No ES = Study met review standards, but did not report sufficient information to calculate a usable effect size

Cells with multiple entries indicate more than one sample

[a]Study used test anxiety level as a selection criterion

Table 5 shows intervention characteristics. The majority of studies used some form of intervention that could be broadly classified into one of five categories: relaxation, expressive writing, support messages (emotional, cognitive, or motivational), music (with or without lyrics), and cognitive reappraisal. Additionally, one study (Edwards, 2012) used exercise, and four studies included multiple component interventions, such as expressive writing and study skills training (Husni, 2006) or test taking skills, relaxation, and cognitive restructuring (Bishop, 2007). The implementer qualifications listed in Table 5 indicate if the implementer of the intervention had any special qualifications. The majority of the interventions were conducted by the teacher/researcher (e.g., reading a script, assigning a writing activity, playing an audio CD), though some required specialized software.

Table 5. *Intervention Characteristics*

Study Author(s) and Year	Study Inclusion	Type of test	Intervention Type	Comparison Group	Delivery format	Delivery mode	Actual frequency of intervention	Actual number of sessions	Time per session (in minutes)	Implementer qualifications	Follow up period (in weeks)
Bishop, 2007[a]	Main	State test	Test taking skills, Relaxation, Cognitive restructuring	BAU	Group	In person	Weekly	4	60	Licensed therapist	0
Dolton, 2016	Main	Standardized test	Relaxation	Neutral listening activity	Group	In person	Weekly	6	16	Audio CD	0
Evans et al., 2010[a]	Main	Standardized test	Relaxation	BAU	Group orientation, followed by self paced recorded training	In person	One group orientation, followed by using training CD at least once	Self paced	50 minutes in person, 31 minute CD	Licensed therapist	20
Harrison, 2016	Main	Standardized test	Music (with lyrics)	BAU	Group	In person	Once	1	3	Teacher	0
Harrison, 2016	Main	Standardized test	Music (without lyrics)	BAU	Group	In person	Once	1	3	Teacher	0
Haynes, 2003	Main	Classroom exam	Music	BAU	Whole class	In person	Once	1	10	Teacher	0

Table 5. *Intervention Characteristics*

Study Author(s) and Year	Study Inclusion	Type of test	Intervention Type	Comparison Group	Delivery format	Delivery mode	Actual frequency of intervention	Actual number of sessions	Time per session (in minutes)	Implementer qualifications	Follow up period (in weeks)
Hines, 2011	Main	State test	Expressive writing	Neutral writing	Whole class	In person	Daily	3	15	Teacher	0
Huang and Mayer, 2016	Main	Lab-based test	Support messages, Relaxation	BAU	Individual	In person	Once	1	8	Software	0
Husni, 2006	Main	Institutional exam	Expressive writing, Study skills training	BAU	Whole class	In person	Daily	40	whole class period	Teacher	0
Jamieson et al., 2016	Main	Classroom exam	Cognitive reappraisal	"ignore stress" message	Individual	In person	Once	1	6.5	Teacher	0
Namwamba, 2013	Main	Lab-based test	Music	BAU	Group	In person	Once	1	Duration of test	Audio CD	0
Nelson and Knight, 2010	Main	Classroom exam	Expressive writing	Neutral writing	Individual	In person	Once	1	NS	Teacher	0
Spielberger, 2015	Main	Classroom exam	Expressive writing	Neutral writing	Individual	In person	Once	1	10	Teacher	0
Brady et al., 2017	Supp	Classroom exam	Cognitive reappraisal	BAU	Individual	Online	Once	1	NS	Teacher	0
Falcon, 2017	Supp	State test	Music	BAU	Whole class	In person	every three weeks	8	45	Teacher	0

Table 5. *Intervention Characteristics*

Study Author(s) and Year	Study Inclusion	Type of test	Intervention Type	Comparison Group	Delivery format	Delivery mode	Actual frequency of intervention	Actual number of sessions	Time per session (in minutes)	Implementer qualifications	Follow up period (in weeks)
Goldenberg et al., 2013	Supp	Classroom exam	Music	BAU	Group (study sessions) and whole class (test)	In person	Not specified	Self paced	NS	Music	0
Im, 2013	Supp	Lab-based test	Support messages (Emotional)	BAU	Individual	In person	Once	1	NS	Software	0
Im, 2013	Supp	Lab-based test	Support messages (Emotional and cognitive)	BAU	Individual	In person	Once	1	NS	Software	0
Im, 2013	Supp	Lab-based test	Support messages (Cognitive)	BAU	Individual	In person	Once	1	NS	Software	0
Insalaca, 2007	Supp	Classroom exam	Music	BAU	Whole class	In person	Daily	40	5	Teacher	0
Kim et al., 2017[a]	Supp	Assessments developed for the curricular content	Support messages	BAU	Individual	Online	Daily	4	4-5 messages per lesson	Software	0

Table 5. *Intervention Characteristics*

Study Author(s) and Year	Study Inclusion	Type of test	Intervention Type	Comparison Group	Delivery format	Delivery mode	Actual frequency of intervention	Actual number of sessions	Time per session (in minutes)	Implementer qualifications	Follow up period (in weeks)
Park et al., 2014[a]	Supp	Lab-based test	Expressive writing	BAU	Individual	In person	Once	1	7	Researcher	0
Perez, 2005	Supp	State test	Expressive writing	BAU	Whole class	In person	3 times a week	12	5	Teacher	0
Shen, 2009[a]	Supp	Lab-based test	Support messages (Emotional)	BAU	Individual	In person	Once	1	NS	Software	0
Shen, 2009[a]	Supp	Lab-based test	Support messages (Motivational)	BAU	Individual	In person	Once	1	NS	Software	0
Shen, 2009[a]	Supp	Lab-based test	Support messages (Emotional and motivational)	BAU	Individual	In person	Once	1	NS	Software	0
Shobe et al., 2005	Supp	Lab-based test	Relaxation	BAU	Individual	In person	Once	1	"less than three minutes" (p. 40)	Teacher	0
Thompson et al., 2016	Supp	Standardized test	Exercise	BAU	Whole class	In person	Once	1	40	Teacher	0

Table 5. *Intervention Characteristics*

Study Author(s) and Year	Study Inclusion	Type of test	Intervention Type	Comparison Group	Delivery format	Delivery mode	Actual frequency of intervention	Actual number of sessions	Time per session (in minutes)	Implementer qualifications	Follow up period (in weeks)
Blank-Spadoni, 2013	No ES	Classroom exam	Expressive writing	Factual writing prompt	Individual	In person	Once	1	10	Teacher	0
Edwards, 2012	No ES	Classroom exam	Exercise	BAU	Whole class	In person	3/week	24	5-10	Teacher	0
Edwards, 2012	No ES	Classroom exam	Exercise	BAU	Whole class	In person	3/week	24	15-60	Teacher	0
Ganzenmuller, 2007	No ES	Standardized test	Relaxation	BAU	Whole class	In person	More than once a week	19 classes plus 5 practice exams	4	Recording	0
Harris et al., 2019	No ES	Classroom exam	Cognitive reappraisal, Expressive writing	analogous neutral task	Individual (Reappraisal); Whole Class(Expressive Writing)	Online (Reappraisal); In person (Expressive Writing)	Once	4 (Reappraisal); 4(Expressive writing)	15 (Reappraisal), 3 (Expressive writing)	Teacher	0
Harrison, 2016	No ES	Standardized test	Music (with lyrics)	BAU	Group	In person	Once	1	3	Teacher	0

Table 5. *Intervention Characteristics*

Study Author(s) and Year	Study Inclusion	Type of test	Intervention Type	Comparison Group	Delivery format	Delivery mode	Actual frequency of intervention	Actual number of sessions	Time per session (in minutes)	Implementer qualifications	Follow up period (in weeks)
Harrison, 2016	No ES	Standardized test	Music (without lyrics)	BAU	Group	In person	Once	1	3	Teacher	0
Henslee & Klein, 2017[a]	No ES	Lab-based test	Relaxation	Close eyes, but remain awake, for 20 minutes	Individual	In person	Once	1	20	recording of imagery session	0
Insalaca, 2007	No ES	Classroom exam	Music	BAU	Whole class	In person	Daily	40	5	Teacher	0
Jacobs, 2021	No ES	Standardized test	Expressive writing	BAU	Individual	Online	Once	1	10	Software	0
Jacobs, 2021	No ES	Standardized test	Cognitive reappraisal	BAU	Individual	Online	Once	1	10	Software	0
Miller et al., 2006[a]	No ES	State test	Relaxation	BAU	Group	In person	five times over the course of half a school year	5	31	school counselor , audio CD	1
Miller et al., 2007[a]	No ES	State test	Relaxation	BAU	Group	In person	Weekly	3	not specified	Not specified	1.5

Table 5. *Intervention Characteristics*

Study Author(s) and Year	Study Inclusion	Type of test	Intervention Type	Comparison Group	Delivery format	Delivery mode	Actual frequency of intervention	Actual number of sessions	Time per session (in minutes)	Implementer qualifications	Follow up period (in weeks)
Sefton, 2014	No ES	Classroom exam	Expressive writing	Objective writing	Whole class	In person	prior to each unit test	7	10	Teacher	0

BAU = 'No Treatment/Business as usual'

NS = Not Specified

Supp = Supplemental Studies

No ES = Study met review standards, but did not report sufficient information to calculate a usable effect size

[a]Study used test anxiety level as a selection criterion

Table 6 shows research design and study quality indicators. Studies labeled “RCT (High Attrition)” were RCT studies that included sufficient information to calculate overall and differential attrition and had attrition that was above the “cautious” boundary set by WWC review standards. These studies were treated as QEDs in the analysis, which also means they met the requirements to be included in the analysis as a QED. Studies included in the supplementary analysis labeled “RCT (Unknown Attrition)” did not provide sufficient information to calculate overall or differential attrition. Most of the studies included in the analysis measured the outcome as an immediate follow up or after a few weeks and only one study (Evans, Ramsey, & Driscoll, 2010) measured long-term test performance (approximately 20 weeks after the intervention). Few studies provided study-specific reliability information for the test anxiety measure and none of the studies provided study-specific reliability information for the test performance measure.

Table 6. *Research design and study quality*

Study Author(s) and Year	Study Inclusion	Publication Status	RCT/QED	Test anxiety measure	Is the test anxiety measure an established measure of test anxiety?	Study-specific reliability for the test anxiety measure	Type of test
Bishop, 2007	Main	Dissertation	RCT	Experimenter developed checklist	No	NS	State test
Dolton, 2016	Main	Dissertation	QED	Cognitive Test Anxiety Scale	Yes, unaltered	0.92	Standardized test
Evans et al., 2010	Main	Non-peer reviewed	RCT	Westside Test Anxiety Scale	Yes, unaltered	NS	Standardized test
Harrison, 2016	Main	Dissertation	RCT	Researcher created survey	No	.70 (pretest), .65 (posttest)	Standardized test
Haynes, 2003	Main	Dissertation	RCT	Math Anxiety Rating Scale	Yes, unaltered	0.96	Classroom exam
Hines, 2011	Main	Dissertation	QED	Math Anxiety Rating Scale	Yes, unaltered	NS	State test
Huang and Mayer, 2016	Main	Journal	RCT	Researcher created open ended questions	No	NS	Lab-based test
Husni, 2006	Main	Dissertation	QED	Math Anxiety Rating Scale	Yes, unaltered	NS	Institutional exam
Jamieson et al., 2016	Main	Journal	RCT (High Attrition)[a]	Abbreviated Math Anxiety Scale	Yes, unaltered	learning subscale: .79, evaluation subscale: .84	Classroom exam
Namwamba, 2013	Main	Dissertation	RCT	Math Anxiety Rating Scale	Yes, unaltered	NS	Lab-based test

Table 6. *Research design and study quality*

Study Author(s) and Year	Study Inclusion	Publication Status	RCT/QED	Test anxiety measure	Is the test anxiety measure an established measure of test anxiety?	Study-specific reliability for the test anxiety measure	Type of test
Nelson and Knight, 2010	Main	Journal	RCT	Spielberger Test Anxiety Inventory	Yes, but has been modified	0.95	Classroom exam
Spielberger, 2015	Main	Dissertation	RCT	Spielberger Test Anxiety Inventory	Yes, unaltered	NS	Classroom exam
Brady et al., 2017	Supp	Journal	RCT (Unknown Attrition)	Researcher created question	No	NS	Classroom exam
Falcon, 2017	Supp	Dissertation	RCT (Unknown Attrition)	Test Anxiety Questionnaire (Nist and Diehl, 1990)	Yes, unaltered	NS	State test
Goldenberg et al., 2013	Supp	Journal	RCT (Unknown Attrition)	Spielberger Test Anxiety Inventory 8-item version modified by Hong (1988)	Yes, unaltered	0.88	Classroom exam
Im, 2013	Supp	Dissertation	RCT (Unknown Attrition)	Mathematics Anxiety Questionnaire	Yes, unaltered	NS	Lab-based test
Insalaca, 2007	Supp	Dissertation	RCT (Unknown Attrition)	Math Anxiety Rating Scale	Yes, unaltered	NS	Classroom exam
Kim et al., 2017	Supp	Journal	RCT (Unknown Attrition)	Revised Mathematics Anxiety Rating Scale	Yes, unaltered	.91 (pretest), .94 (posttest)	Assessments developed for the curricular content

Table 6. *Research design and study quality*

Study Author(s) and Year	Study Inclusion	Publication Status	RCT/QED	Test anxiety measure	Is the test anxiety measure an established measure of test anxiety?	Study-specific reliability for the test anxiety measure	Type of test
Park et al., 2014	Supp	Journal	RCT (Unknown Attrition)	Short Math Anxiety Rating Scale	Yes, unaltered	NS	Lab-based test
Perez, 2005	Supp	Dissertation	RCT (Unknown Attrition)	Math Anxiety Rating Scale	Yes, unaltered	.98 (pretest), .99 (posttest)	State test
Shen, 2009	Supp	Dissertation	RCT (Unknown Attrition)	Math Anxiety Rating Scale, Math Anxiety Questionnaire	Yes, unaltered	.86 (MAQ affective reactions), .76 (MAQ cognitive worrying), .89 MAS	Lab-based test
Shobe et al., 2005	Supp	Journal	RCT (Unknown Attrition)	Spielberger Test Anxiety Inventory	Yes, unaltered	NS	Lab-based test
Thompson et al., 2016	Supp	Journal	RCT (Unknown Attrition)	Children's Test Anxiety Scale	Yes, unaltered	NS	Standardized test
Blank-Spadoni, 2013	No ES	Dissertation	RCT (Unknown Attrition)	Costello-Comrey Anxiety Inventory	Yes, unaltered	0.87	Classroom exam
Edwards, 2012	No ES	Dissertation	RCT (High Attrition)[a]	None	No	NS	Classroom exam

Table 6. *Research design and study quality*

Study Author(s) and Year	Study Inclusion	Publication Status	RCT/QED	Test anxiety measure	Is the test anxiety measure an established measure of test anxiety?	Study-specific reliability for the test anxiety measure	Type of test
Edwards, 2012	No ES	Dissertation	RCT	None	No	NS	Classroom exam
Ganzenmuller , 2007	No ES	Dissertation	QED	Reaction to Tests Questionnaire	Yes, unaltered	.83 to .93 across four subscales	Standardized test
Harris et al., 2019	No ES	Journal	RCT (Unknown Attrition)	Cognitive Test Anxiety Scale	Yes, unaltered	NS	Classroom exam
Harrison, 2016	No ES	Dissertation	RCT (Unknown Attrition)	Researcher created survey	No	.70 (pretest), .65 (posttest)	Standardized test
Henslee & Klein, 2017	No ES	Journal	RCT (Unknown Attrition)	Math Anxiety Rating Scale	Yes, unaltered	NS	Lab-based test
Insalaca, 2007	No ES	Dissertation	RCT (Unknown Attrition)	Math Anxiety Rating Scale	Yes, unaltered	NS	Classroom exam
Jacobs, 2021	No ES	Dissertation	RCT (Unknown Attrition)	Spielberger Test Anxiety Inventory (short form)	Yes, but has been modified	NS	Standardized test
Miller et al., 2006	No ES	Non-peer reviewed	RCT (High Attrition)[a]	Westside Test Anxiety Scale	Yes, unaltered	NS	State test

Table 6. *Research design and study quality*

Study Author(s) and Year	Study Inclusion	Publication Status	RCT/QED	Test anxiety measure	Is the test anxiety measure an established measure of test anxiety?	Study-specific reliability for the test anxiety measure	Type of test
Miller et al., 2007	No ES	Non-peer reviewed	RCT (High Attrition)[a]	Westside Test Anxiety Scale	Yes, unaltered	NS	State test
Sefton, 2014	No ES	Dissertation	RCT (High Attrition)[a]	Cognitive Test Anxiety Scale	Yes, unaltered	NS	Classroom exam

[a]Study was an RCT with high attrition or attrition could not be calculated. Treated as QED in this analysis.

NS = Not Specified

Supp = Supplemental Studies

No ES = Study met review standards, but did not report sufficient information to calculate a usable effect size

Cells with multiple entries indicate more than one outcome

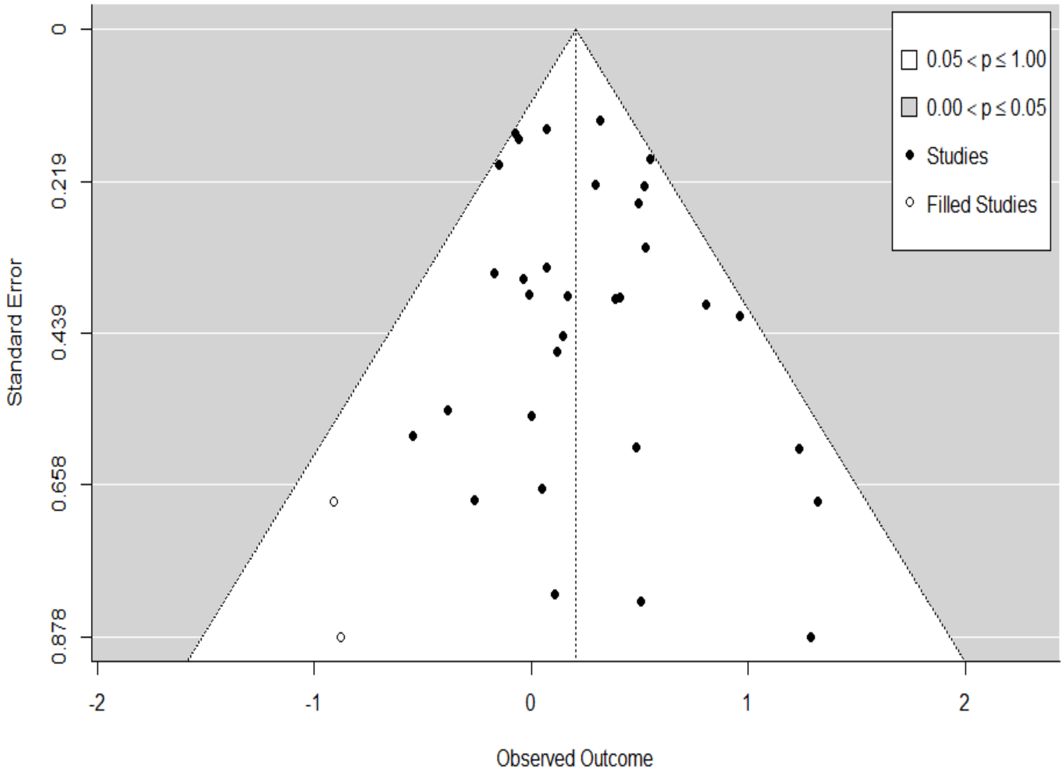
All studies included in the analysis had used test anxiety and test performance measures that were face valid

No studies reported study-specific reliability for the test performance measure except for: Huang and Mayer, 2016

Publication and Reporting Bias

To help reduce the risk of publication bias, both published and unpublished studies were included in the analysis. As seen earlier in Table 6, the majority (9 of 12) studies included in the main analysis were unpublished dissertations or online submissions to ERIC. A moderator analysis using publication status (published vs. unpublished) found no significant difference between studies based on publication status (p = .81); the mean effect size of unpublished studies was .03 standard deviations smaller than those reported in published studies. Similarly, Begg and Mazumdar's rank correlation test (τ = 0.11, p = .38) and Egger's regression test (z = 1.10, p = .27) were not statistically significant. Finally, a trim and fill analysis (Figure 3) was used to assess the meta-analytic dataset for the possible presence of publication bias. In the resulting adjusted funnel plot (Figure 3), the solid dots represent the point estimates for the studies included in the analysis. The hollow dots are the symmetrically filled estimates, that is, imputed mirror images of the observed estimates that the trim and fill algorithm identified as causing funnel plot asymmetry. As can be seen in the funnel plot, the two filled estimates are to the left of the mean effect size. The trim and fill estimate of the overall effect size was 0.21, which was not substantially different from the effect size based only on the observed outcomes (g = 0.22).

Figure 2. *Funnel plot with trim and fill*



Assessment of Internal Validity

A moderation test was conducted to determine if effect sizes varied by study design. The mean effect sizes reported in RCTs and QED did not significantly differ (*p* = .84). RCTs had an effect size only .02 standard deviations higher that QEDs. Since effect sizes were similar by study design, there was no need to control for study design in later moderation tests as planned. An additional moderation test comparing the meta-analytic mean effect size for RCTs with known attrition to RCTs with unknown attrition was not significantly different (*p* = .60). Unknown attrition RCTs had an effect size .07 standard deviations lower than RCTs with known attrition. With no statistical or substantive difference between the effect sizes in either group, I decided to include the 11 RCTs with unknown attrition in all remaining analyses. Doing so increased the number of studies available in the analysis and increased the statistical power of the moderation tests. The slightly lower effect size noted in the RCTs with unknown attrition should also lead to somewhat more conservative estimates in later analyses.

There was evidence of outcome reporting bias in multiple studies. Outcome reporting bias is a threat to the validity of any effort to synthesize the research in an area because it can lead to inaccurate, and typically inflated, estimates of the effect of interventions (Pigott, Valentine, Polanin, Williams, & Canada, 2013). Typically, effect sizes or means and standard deviations were only reported when the statistical test reported a significant result. In such cases, it was also common that incomplete information for the statistical test was reported (e.g., only the overall *F* statistic being reported for an ANCOVA), meaning it was not possible to use what limited information was available to calculate an effect size. The most egregious case was found in Harrison (2016). Of the eight outcomes reported that were eligible for this analysis, only the two outcomes with statistically significant results included tables of means and standard deviations. Of these, the results from the pilot study were a clear outlier, with an effect size of 2.41 for the math test in the music with lyrics intervention. Given the unusual results, I decided to remove the pilot study from the analysis. The second, larger, study, labeled the "actual" study in Harrison remained in my meta-analytic data set.

Roughly a third of studies (8 out of 23) used cluster assignment; however, none of these studies used statistical methods to control for clustering, leading to a violation of the assumption of nonindependence. Violation of the assumption of nonindependence of observations is particularly problematic, as it leads to underestimating the variability in the outcome (McCoach & Adelson, 2010). To account for clustering, I calculated the standard error of the effect size for the studies that used cluster assignment using the formula provided in the WWC *Procedures Handbook* version 4.1 (Formula E.5.2; U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse, 2020b). Since none of the studies that used cluster assignment reported the intraclass correlation, I used the WWC default of .20 for academic achievement outcomes (see also Hedges & Hedberg, 2007).

Meta-Analytic Results

Table 7 shows effect size information for each outcome. Where a study reported multiple outcomes for the same sample, the outcomes can be distinguished by the "Outcome" and "Intervention Type" columns. Some studies examined the effect of multiple interventions and compared the results of these interventions to a common control group. Both of these situations (multiple outcomes and multiple intervention conditions that share a control group) mean that the effect sizes are not entirely independent. Originally, I had planned to use robust variance estimation to account for non-independence, however, there was not enough data (studies, and non-independent effects within studies) to use this approach. Instead, to address the problem of multiple interventions compared to a common control group, I adjusted the standard error of the effect size by dividing the comparison sample size by the number of interventions reported (i.e., if a study contained two interventions and a shared comparison group, the comparison sample size was divided by 2; Higgins, Deeks, Altman, 2011, Chapter 16.5.4). For studies that reported multiple outcomes for the same sample, I used the "aggregate" function in the "metafor" package to calculate a study average effect size and standard error, the latter of which is adjusted to account for the fact that the measures are not perfectly correlated. To use this method, I had to impute a typical correlation between measures; I chose +.60, but results are extremely similar across all reasonable values of this correlation.

Table 7. *Effect Size Information*

Study Author(s) and Year	Study Inclusion	Outcome	Intervention Type	Sample label	*g*	*SEg*	Analytic sample overall N	Analytic sample overall cluster N
Bishop, 2007	Main	MontCAS - Reading	Test taking skills, Relaxation, Cognitive restructuring	Full Sample	0.84	0.38	30	2
Bishop, 2007	Main	MontCAS - Mathmatics	Test taking skills, Relaxation, Cognitive restructuring	Full Sample	0.15	0.37	30	2
Dolton, 2016	Main	WRAT-4 - Word-Reading subscale	Relaxation	Full Sample	-0.20	0.43	22	-
Dolton, 2016	Main	WRAT-4 - Sentence-Comprehension subscale	Relaxation	Full Sample	0.18	0.43	22	-
Dolton, 2016	Main	WRAT-4 - Spelling subscale	Relaxation	Full Sample	0.28	0.44	22	-
Dolton, 2016	Main	WRAT-4 - Math-Computation subscale	Relaxation	Full Sample	0.02	0.43	22	-
Evans et al., 2010	Main	HESI Exit Exam	Relaxation	Full Sample	0.53	0.31	42	-
Harrison, 2016	Main	ACT Reading practice test	Music (with lyrics)	Music (with lyrics)	-0.54	0.59	18	-
Harrison, 2016	Main	ACT Reading practice test	Music (without lyrics)	Music (without lyrics)	-0.38	0.55	22	-

Table 7. *Effect Size Information*

Study Author(s) and Year	Study Inclusion	Outcome	Intervention Type	Sample label	*g*	*SEg*	Analytic sample overall N	Analytic sample overall cluster N
Haynes, 2003	Main	College Algebra Exam 1	Music	Full Sample	0.12	0.16	160	4
Hines, 2011	Main	Standard of Learning Mathematics Test (Practice version)	Expressive writing	Full Sample	0.15	0.21	93	5
Huang and Mayer, 2016	Main	Retention test	Support messages, Relaxation	Full Sample	0.62	0.28	54	-
Huang and Mayer, 2016	Main	Transfer test	Support messages, Relaxation	Full Sample	0.27	0.27	54	-
Huang and Mayer, 2016	Main	Practice test	Support messages, Relaxation	Full Sample	0.70	0.28	54	-
Husni, 2006	Main	College Placement Exam	Expressive writing, Study skills training	Full Sample	0.05	0.26	60	2
Jamieson et al., 2016	Main	Math test	Cognitive reappraisal	Full Sample	0.30	0.22	81	-
Namwamba, 2013	Main	"Algebra Ability Instrument"	Music	10 db relative volume	0.11	0.82	12	-
Namwamba, 2013	Main	"Algebra Ability Instrument"	Music	20 db relative volume	0.51	0.83	12	-
Namwamba, 2013	Main	"Algebra Ability Instrument"	Music	30 db relative volume	1.29	0.88	12	-

Table 7. *Effect Size Information*

Study Author(s) and Year	Study Inclusion	Outcome	Intervention Type	Sample label	*g*	*SEg*	Analytic sample overall N	Analytic sample overall cluster N
Nelson and Knight, 2010	Main	15 item pop quiz	Expressive writing	Full Sample	0.55	0.19	118	-
Spielberger, 2015	Main	Teacher designed test	Expressive writing	Full Sample	-0.15	0.19	106	-
Brady et al., 2017	Supp	Exam 1 grade	Cognitive reappraisal	Upper-level students	0.07	0.14	194	-
Brady et al., 2017	Supp	Exam 1 grade	Cognitive reappraisal	First-year students	0.32	0.13	237	-
Falcon, 2017	Supp	Reading Comprehension	Music	7th graders	-0.26	0.32	42	2
Falcon, 2017	Supp	Reading Comprehension	Music	8th graders	1.32	0.30	53	2
Goldenberg et al., 2013	Supp	Second course exam	Music	Full Sample	-0.07	0.15	176	-
Im, 2013	Supp	Math problem solving	Support messages (Emotional)	Emotional support	0.17	0.38	41	-
Im, 2013	Supp	Math problem solving	Support messages (Emotional and cognitive)	Emotional and cognitive support	0.80	0.40	41	-
Im, 2013	Supp	Math problem solving	Support messages (Cognitive)	Cognitive support	-0.01	0.38	42	-
Insalaca, 2007	Supp	Final exam	Music	High School geometry	0.00	0.23	80	3

Table 7. *Effect Size Information*

Study Author(s) and Year	Study Inclusion	Outcome	Intervention Type	Sample label	*g*	*SEg*	Analytic sample overall N	Analytic sample overall cluster N
Kim et al., 2017	Supp	Module-based learning assessments	Support messages	High anxiety group	-0.04	0.36	32	-
Park et al., 2014	Supp	Composite z-score (Average of error rates & reaction times)	Expressive writing	High math anxiety	0.78	0.31	44	-
Park et al., 2014	Supp	Math error rates (high demand problems)	Expressive writing	High math anxiety	0.41	0.30	42	-
Park et al., 2014	Supp	Math error rates (low demand problems)	Expressive writing	High math anxiety	0.33	0.30	44	-
Perez, 2005	Supp	Texas Higher Education Assessment (THEA) practice test	Expressive writing	Full Sample	-0.17	0.18	123	8
Shen, 2009	Supp	Math test	Support messages (Emotional)	Emotional Support	0.96	0.41	50	-
Shen, 2009	Supp	Math test	Support messages (Motivational)	Motivational Support	0.39	0.39	55	-
Shen, 2009	Supp	Math test	Support messages (Emotional and motivational)	Emotional and motivational support	0.41	0.39	56	-
Shobe et al., 2005	Supp	Easy math test	Relaxation	Full Sample	0.88	0.66	10	-

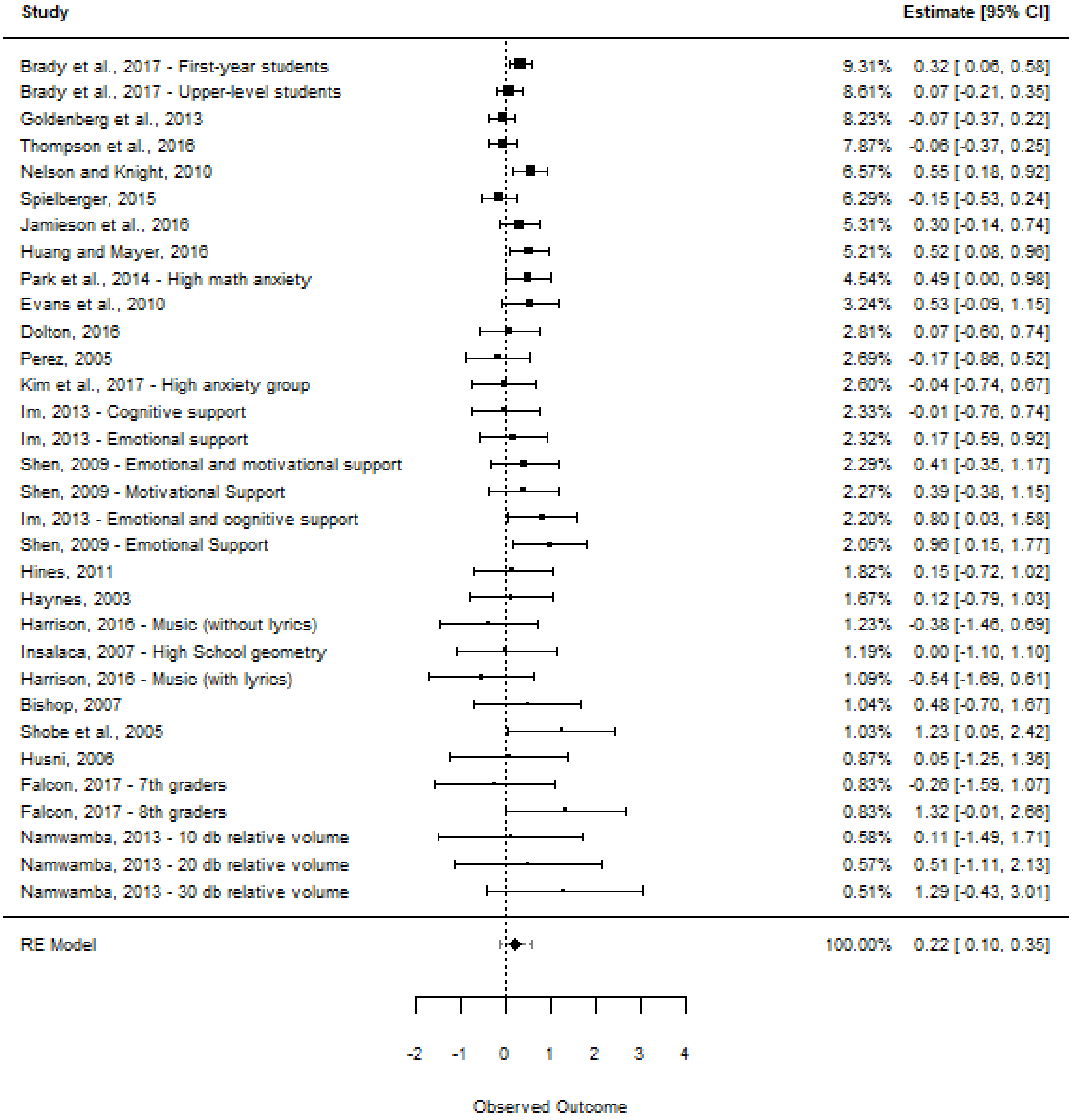
Table 7. *Effect Size Information*

Study Author(s) and Year	Study Inclusion	Outcome	Intervention Type	Sample label	*g*	*SEg*	Analytic sample overall N	Analytic sample overall cluster N
Shobe et al., 2005	Supp	Difficult math test	Relaxation	Full Sample	1.81	0.75	10	-
Thompson et al., 2016	Supp	NWEA MAP (Math)	Exercise	Full Sample	-0.02	0.07	791	29
Thompson et al., 2016	Supp	NWEA MAP (Reading)	Exercise	Full Sample	-0.10	0.08	709	26

Supp = Supplemental Studies

A graphical overview of the studies that contributed data to this analysis can be seen in the forest plot in Figure 3. Study effect sizes reported here are aggregated to the independent sample level (i.e., if a sample has two outcomes, the effect size reported in Figure 3 is the weighted mean). Studies are ordered by sample variance, such that studies with the smallest variances, and by extension weighted the most in the analysis, are shown at the top of the plot, while studies with the largest variances are at the bottom. The squares represent the point estimates of the study's effect size, plotted on the x-axis. The size of the square is related to the inverse of the standard error of the effect size, the larger the point, the smaller the standard error. The wings extending from the point estimates represent the 95% confidence interval for the effect size.

Figure 3. *Forest plot*



The overall meta-analytic mean effect size for test anxiety interventions on test performance was a standardized mean difference of 0.22 ($p < .001$), which equates to a Cohens U_3 of .59, indicating that 59% of intervention students would be expected to score above the mean of the comparison group. Alternatively, the size of this effect can be interpreted, using the Common Language Effect Size, as indicating that a student in the intervention group would have a 56% probability of scoring higher than a student in the comparison group (*CLES* = .56). Of course, 50% in this context would be the expected value of both translations if the null hypothesis is true ($\delta = 0$).

The test of homogeneity was not statistically significant, $Q(31) = 37.92$, $p = .18$, meaning I cannot reject the null hypothesis that the studies in the analysis are estimating the same population parameter. I^2 was 24%, which indicates a non-trivial proportion of the variance in the effect sizes is due to true variability in effect sizes, as opposed to random sampling error. As can be seen in the above figure, the diamond at the bottom of the plot represents the mean effect size and the wings extending from it is the 95% prediction interval for the mean effect size. The prediction interval provides an estimate of the distribution of true effect sizes. If another study, with characteristics similar to those observed in my meta-analysis, of the effect of a test anxiety intervention on test performance were to be conducted, the effect size of this hypothetical study would be expected to be between -0.12 and 0.56. This is a substantial range of possible effect sizes, providing additional evidence of the heterogeneity, or variability among effect sizes that cannot be attributed to sampling error, within these studies.

Moderation analyses

To get a clearer view of the effect of test anxiety interventions on test performance I conducted three moderation tests: type of intervention/therapeutic approach, academic level, and test subject.

Intervention type

A moderation test for effect of intervention type was conducted. The multicomponent intervention studies were not included in this test due to the degree of heterogeneity between the types of interventions provided. Similarly, the exercise intervention was omitted from this moderation analysis because there was only one study that used this type of intervention. As seen in Table 8, of the intervention types tested, relaxation interventions had the largest effect size, while music interventions had essentially no effect. Despite the wide spread of effect sizes, an omnibus test of intervention types was not statistically significant, $Q_{model}(4) = 3.29, p = .51$.

Table 8. *Effect Size by Intervention Type*

Intervention Type	# of Studies	# of Effect Sizes	*g*	*SE*
Relaxation	3	3	0.46	0.25
Support messages	3	7	0.36	0.15
Cognitive reappraisal	2	3	0.22	0.09
Expressive writing	5	5	0.21	0.17
Music	6	10	-0.01	0.12

Note: Pairwise comparisons found no statistically significant difference between intervention types

Academic level

To test for the effect of test anxiety interventions at different academic levels, a moderation test was conducted. Elementary samples had the smallest effect size, while secondary and postsecondary school samples had the same effect size, though an omnibus test of academic levels was not statistically significant,$Q_{model}(2) = 1.81, p = .41$.

Table 9. *Effect Size by Academic Level*

Academic Level	# of Studies	# of Effect Sizes	*g*	*SE*
Elementary	2	2	-0.03	0.14
Secondary	7	12	0.25	0.13
Postsecondary	13	16	0.25	0.08

Note: Pairwise comparisons found no statistically significant difference between academic levels

Subject of test

The majority of studies used either math or English language arts tests as their outcomes, though there were a few that used psychology or nursing tests, as well. For the purposes of this moderation test, I am only testing the difference between math and English language arts subjects because there were not enough outcomes for the other tests to include them. Table 10 shows effect size information by test subject. Despite the noticeable difference between meta-analytic effect sizes, an omnibus test of academic levels was not statistically significant, $Q_{\text{model}}(1) = 1.24$, $p = .26$.

Table 10. *Effect Size by Test Subject*

Subject of test	# of Studies	# of Effect Sizes	*g*	*SE*
ELA	5	7	-0.02	0.14
Math	16	22	0.28	0.08

Note: Pairwise comparisons found no statistically significant difference between subject types

DISCUSSION

This systematic review and meta-analysis included 42 effect sizes nested in 23 studies of interventions focused on test anxiety in which test performance was a measured outcome. Overall, there was a small effect of test anxiety interventions on test performance. There was a sizeable amount of heterogeneity in the overall analysis and subsequent moderation analyses found a wide spread of effect sizes based on intervention type, academic level of the sample, and subject of test. None of these differences in effect sizes were statistically significant, though. Given the small number of effect sizes present in some of the intervention groups and academic levels, the moderation tests were underpowered and, given their univariate nature, they were likely at least somewhat confounded with other study characteristics. For example, while studies that used English language arts tests drew their samples evenly from elementary and secondary school contexts, studies that used math tests drew almost exclusively from high school and postsecondary contexts. Similarly, most of the interventions used in studies with ELA outcomes were relaxation or music, whereas there was a greater diversity of interventions among the studies with math outcomes.

One positive trend noted in the research is the number of studies that attempted whole classroom interventions. While such interventions introduce their own complications, they have wider applicability compared to single student or small group interventions. Interventions designed to treat individual students introduce issues related to equitable access to treatment as, at the grade school level, they would likely require a student to have 504 plan or IEP in place before schools would be able to provide services. Pursuing a formal diagnosis of test anxiety, especially in light of the fact that there are no specific test anxiety criteria in the DSM-5, would require time and family resources not all students will have access to. Given the prevalence of test anxiety, whole classroom or school-wide approaches point the way toward universal prevention strategies as a means of impacting the quality of students' lives and improving school metrics beyond those measured by accountability regimes. Additional research in this direction is warranted.

Some general notes regarding the quality of the data included in this synthesis are worth highlighting. First, there was some evidence of publication bias, as was shown in the funnel plot in Figure 2. The trim and fill procedure identified 2 estimates that contributed to a lack of symmetry in the funnel plot. Adjusting for the presence of this asymmetry, though, results in an estimated effect that is largely the same. While there were too few studies in the subsets of estimates included in the moderation analyses, it would be reasonable to expect similar findings at that level as well. Next, there is the issue of data reporting. There were quite a few studies that were generally poorly reported. As mentioned in the Results section, underreporting of outcomes was common. Approximately a third of outcomes that met standards for this review, twenty-three outcomes from 11 studies, failed to report sufficient information to calculate effect sizes. It should be noted, of course, that a nonsignificant *p*-value does not necessarily mean that the effect size can be assumed to be zero. It may simply mean that the sample size was not large enough to detect the effect that may have been present. In other words, the nonsignificant *p*-value may suggest a null effect size or that the study did not have sufficient statistical power to detect the actual effect. Either way these missing outcomes represent a loss for this analysis, as even small studies reporting small effects could have provided additional data and reduced the risk of bias in this analysis. As can be seen by the overall sample sizes in Table 7, very few studies included in the current analysis appear to be adequately powered to detect an effect size of the size of the meta-analytic mean found in the current analysis.

Underreporting of basic sample information, such as race/ethnicity, age, and region the sample was drawn from was also common. This can be problematic as school and district leadership often look for, or place a greater emphasis on, interventions or programs that show evidence of being successful in schools which are demographically similar to their own. The number of RCT studies labeled in this analysis as “Unknown Attrition” is another example of a lack of thorough reporting. Nearly half of all studies included in this analysis did not report baseline sample sizes, either overall or by intervention/comparison group, or the study was not clear that there was no attrition. Often, when baseline sample sizes were reported, the sources of sample loss from baseline to outcome were not clearly reported.

A conceptual issue with the music interventions, which in a way is related to the issue of underreporting in general, is that the majority of these studies were vague regarding the type of music used. Often only stating that “classical music” was used. Goldenberg et al. (2013) specified music by Mozart and Harrison used a self-developed song designed “to encourage students to recall information and increase their motivation” (2016, p. 74), though Harrison did not provide any additional details regarding this self-developed song. The ambiguity is relevant here because both Mozart’s “Eine kleine Nachtmusik” and Tchaikovsky’s “1812 Overture” would be classified as “classical,” but the effect of listening to these at study or test time would presumably differ. Additionally, only Namwamba (2013) explicitly stated the volume of the music in the intervention, which was shown in their study to have an impact on performance.

Overall, the statistical methods used in most studies included in this analysis were simple. The most common statistical tests used were the *t*-test and ANOVA. While simple statistical methods are not necessarily a problem when paired with rigorous research methods, the research methods (i.e., sample screening, group assignment, etc.) were similarly simple. Given that most of the studies did not include a test anxiety level criterion for inclusion in the study, it would have been reasonable to include test anxiety as a covariate in an ANCOVA or regression equation, especially since all studies in this analysis collected baseline test anxiety data, but none controlled for it. Alternatively, results could have been provided for the subset of the sample that scored high on test anxiety, but only Haynes (2003) did so. Also, while it was common for studies in this analysis to use cluster assignment into condition (whole classrooms being assigned to condition), none of the studies used statistical methods that accounted for clustering. Not accounting for the nonindependence of observations results in underestimating the variability in the outcome (McCoach & Adelson, 2010). Though, it should be noted again that many studies in this area did not have large enough sample sizes to utilize more sophisticated statistical methods (i.e., multilevel modeling). I calculated standard errors that were adjusted for the effect of clustering, which resulted in substantially greater variability among the cluster assignment studies than the unadjusted standard errors would suggest.

Concerning the interventions used in the studies in this analysis, there were many studies that relied on very “light touch” interventions, such as listening to music for three minutes prior to a test (Harrison, 2016), a short relaxation exercise prior to a test (Shobe et al., 2005), or a single cognitive reframing email prior to the exam asserting that feeling stress/anxiety can help test performance (Brady et al., 2017). These studies, in particular, seemed to lack a thoughtful consideration of the root causes of test anxiety and instead attempted to identify simple or quick “band-aid” solutions to the complicated problems of test anxiety and test performance. This issue is all the more problematic in light of Theobald, Breitwieser, and Brod’s (2022) recent study, which suggested that the effect of test anxiety on test performance may be greatest during the encoding phase of memory rather than retrieval. If this is the case, then it is not surprising that interventions that take place right before the test have only a small effect. While it is understandable to seek interventions that require minimal time or resources to implement, there is a need for interventions that can be utilized during learning (encoding), not just at testing (retrieval). Future researchers would do well to consider developing a logic model, or directed acyclic graph, while developing their intervention plans.

Limitations

There was a noticeable amount of unexplained heterogeneity among the studies. If I could have obtained more studies, it might have been possible to plan additional moderation tests, particularly multivariate moderation tests, to continue to explore the sources of variation in the research base. One method that I could have used to increase the number of studies eligible for this analysis would have been to include studies from other countries, possibly just in English-speaking countries or extended to OECD countries. This would have, of course, required additional tests to determine if effect sizes varied based on country/region, but the additional studies might have provided sufficient statistical power to identify significant sources of heterogeneity. I chose to limit studies to those that took place in the U.S. out of a desire to maintain applicability in the U.S. school context, but a future analysis could examine to what degree the decision I made on theoretical grounds stands up empirically. Additionally, I noted some interventions were more commonly tested in studies that occurred outside the United States, than within the U.S. and it may have been interesting to synthesize these interventions. For example, aromatherapy interventions were far more common in international studies. None of the American aromatherapy studies met the eligibility requirements or quality standards to be included in this analysis.

A design decision that I made that likely had an impact on the number of studies that could be included in the analysis was the requirement that studies be made publicly available in the past 20 years. This decision was made for theoretical concerns related to the changes in the testing environment in the past two decades, and is in keeping with WWC study review standards, but including older studies, and modeling the effect of publication date, might have opened up additional opportunities.

Next steps

This analysis specifically examined the effect of test anxiety interventions on test performance. A subset of the studies included in the analysis also included test anxiety post-intervention. Of interest from a theoretical point of view would be to analyze the relationship between post-intervention test anxiety and test performance. If this relationship could be modeled, then there would be some support for relying on the much larger research base of studies that used test anxiety reduction as the sole outcome. This would be helpful for teachers and school districts, as the current analysis, unfortunately, gives little in the way of clear guidance for what test anxiety interventions would be best for improving test performance.

The overall poor quality of the research base found during this analysis should be a call to researchers that there is a need for more high-quality research in this area. The consequences of test anxiety can be high as students' performance is underestimated and academic progress can be cut short. So long as students, schools, and districts are judged based on test scores we need to ensure that those test scores are proper reflections of students' actual capabilities. Additional research, thoroughly reported, using larger sample sizes, which control for baseline test anxiety, and interventions designed to address the root causes of test anxiety are necessary to address the shortcomings identified in this review.

REFERENCES

* References marked with an asterisk indicate studies included in the meta-analysis

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. New York: AERA.

American Psychiatric Association. (2022). *Diagnostic and Statistical Manual of Mental Disorders* (5th ed., Text Revision). American Psychiatric Association, Washington, DC.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50*(4), 1088–1101. doi: 10.2307/2533446

Beggs, C., Shields, D., & Goodin, H. (2011). Using guided reflection to reduce test anxiety in nursing students. *Journal of Holistic Nursing*, *29*(2), 140–147. doi: 10.1177/0898010110393352

Beidel, D., Turner, M., & Trager, K. (1994). Test anxiety and childhood anxiety disorders in African American and White school children. *Journal of Anxiety Disorders, 8*(2), 169-179. doi: 10.1016/0887-6185(94)90014-0

Bellinger, D. B., DeCaro, M. S., & Ralston, P. A. (2015). Mindfulness, anxiety, and high-stakes mathematics performance in the laboratory and classroom. *Consciousness and Cognition, 37,* 123-132. doi: 10.1016/j.concog.2015.09.001

*Bishop, N. S. (2007). *Implementation and assessment of a test anxiety reduction program presented to 10th graders and their subsequent performance on the MontCAS criterion referenced test.* [Doctoral dissertation, University of Montana]

Bodas, J., & Ollendick, T.H. (2005). Test anxiety: A cross-cultural perspective. *Clinical Child and Family Psychology Review, 8,* 65-88. doi: 10.1007/s10567-005-2342-x

Borenstein, M., Hedges, L.V., Higgins, J.P.T., & Rothstein, H.R. (2009). *Introduction to Meta-Analysis.* Chichester, UK: John Wiley & Sons, Ltd.

*Brady, S. T., Hard, B. M., & Gross, J. J. (2017). Reappraising test anxiety increases academic performance of first-year college students. *Journal of Educational Psychology, 110*(3), 395-406. 10.1037/edu0000219

Bramer, W. M., Giustini, D., de Jonge, G. B., Holland, L., & Bekhuis, T. (2016). De-duplication of database search results for systematic reviews in EndNote. *Journal of the Medical Library Association: JMLA, 104*, 240-243. doi: 10.5195/jmla.2016.24

Carter, E. W., Wehby, J., Hughes, C., Johnson, S. M., Plank, D. R., Barton-Arwood, S. M., et al. (2005). Preparing adolescents with high-incidence disabilities for high-stakes testing with strategy instruction. *Preventing School Failure, 49*, 55-62. doi: 10.1080/1045988X.2005.10823218

Cassady, J. C. (2004). The influence of cognitive test anxiety across the learning–testing cycle. *Learning and Instruction*, *14*(6), 569-592. doi: 10.1016/j.learninstruc.2004.09.002

Cassady, J.C. (2009). Test anxiety: Contemporary theories and implications for learning. In J.C. Cassady (Ed.), *Anxiety in schools: The causes, consequences, and solutions for academic anxieties* (119-136). New York: Peter Lang Publishing.

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

Chapell, M. S., Blanding, Z. B., Silverstein, M. E., Takahashi, M., Newman, B., Gubi, A., & McCann, N. (2005). Test anxiety and academic performance in undergraduate and graduate students. *Journal of Educational Psychology, 97*, 268. doi: 10.1037/0022-0663.97.2.268

Caviola, S., Toffalini, E., Giofrè, D., Ruiz, J. M., Szűcs, D., & Mammarella, I. C. (2022). Math performance and academic anxiety forms, from sociodemographic to cognitive aspects: A meta-analysis on 906,311 participants. *Educational Psychology Review*, *34*(1), 363-399. doi: 10.1007/s10648-021-09618-5

Covington, M. V. (2000). Goal theory, motivation, and school achievement: An integrative review. *Annual Review of Psychology*, *51*, 171-200. doi: 10.1146/annurev.psych.51.1.171

Davis, H.A., DiStefano, C., & Schutz, P.A. (2008). Identifying patterns of appraising tests in first-year college students: Implications of anxiety and emotion regulation during test taking. *Journal of Educational Psychology, 100,* 942-960. doi: 0.1037/a0013096

DeVellis, R.F. (2012). *Scale development: Theory and applications*. Los Angeles: Sage.

*Dolton, M. G. (2016). *Teaching relaxation techniques to improve achievement and alleviate the anxiety of students with learning disabilities in an independent school.* [Doctoral dissertation, Nova Souheastern University]

Donato, J. M. (2010). *Reducing test anxiety and improving academic performance in fourth grade students: Exploring an intervention.* [Doctoral dissertation, Southern Connecticut State University]

Dowker, A., Sarkar, A., & Looi, C. Y. (2016). Mathematics anxiety: What have we learned in 60 years? *Frontiers in Psychology*, *7*, 508. doi: 10.3389/fpsyg.2016.00508

Duval, S. J., & Tweedie, R. L. (2000a). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*(2), 455–463. doi: https://doi.org/10.1111/j.0006-341x.2000.00455.x

Dreisbach, M. D. (2017). *The Effects of a Classroom Based Yoga Intervention on Test Anxiety, Academic Performance and Attention in Third Grade Students.* [Master's Thesis, University of Akron]

Driscoll, R.; Holt, B. & Hunter, L. (2005). *Accelerated Desensitization and Adaptive Attitudes Interventions and Test Gains with Academic Probation Students* (ED494905). ERIC. http://files.eric.ed.gov/fulltext/ED494905.pdf

Edwards, J. J. (2012). *Strong Body, Strong Mind: The Effects of Implementing Physical Activity within a Mathematics Course for Deployed Sailors*. [Doctoral dissertation, Texas A&M University]

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, *315*(7109), 629-634. doi: 10.1136/bmj.315.7109.629

Ergene, T. (2003). Effective interventions on test anxiety reduction: A meta-analysis. *School Psychology International. 24,* 313-328. doi: 10.1177/01430343030243004

*Evans, G., Ramsey, G., & Driscoll, R. (2010). *Test-Anxiety Program and Test Gains with Nursing Classes* (ED512827). ERIC. http://files.eric.ed.gov/fulltext/ED512827.pdf

Faber, G. (2010). Enhancing orthographic competencies and reducing domain-specific test anxiety: The systematic use of algorithmic and self-instructional task formats in remedial spelling training. *International Journal of Special Education, 25*, 78-88.

*Falcon, E. (2017). *The relationship between background classical music and reading comprehension on seventh and eighth grade students.* [Doctoral dissertation, St. Thomas University]

Fisher, Z., Tipton, E., Zhipeng, H. (2017). robumeta: Robust variance meta-regression. (R package version 2.0) [Computer software]. The Comprehensive R Archive Network. Available from https://CRAN.R-project.org/package=robumeta

Gibson, H.A. (2014). A conceptual view of test anxiety. *Nursing Forum, 49(4),* 267-277. doi: 10.1111/nuf.12069

*Goldenberg, M. A., Floyd, A. H. L. & Moyer, A. (2013). No effect of a brief music intervention on test anxiety and exam scores in college undergraduates. *Journal of Articles in Support of the Null Hypothesis, 10*(1), 1-16.

Han, H., & Lee, J. (2021) The relationship between Korean university students' suicidal ideation and risk factors: a meta-analysis, *International Journal of Adolescence and Youth, 26*(1), 405-420, doi: 10.1080/02673843.2021.1974901

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*(1), 17-27. doi: 10.1111/j.1745-3992.2004.tb00149.x

*Harrison, E Y. (2016). *The relationship between music test attitude and test anxiety: An action research study.* [Doctoral dissertation, Capella University]

Hart, R., Casserly, M., Uzzell, R., Palacios, M., Corcoran, A., & Spurgeon, L. (2015). Student testing in America's great city schools: An inventory and preliminary analysis. *Council of the Great City Schools*. Retrieved from http://www.cgcs.org/cms/lib/DC00001581/Centricity/Domain/87/Testing%20Report.pdf

*Haynes, S. E. (2003). *The effect of background music on the mathematics test anxiety of college algebra students.* [Doctoral dissertation, West Virginia University]

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107-128. doi: 10.2307/1164588

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60-87. doi: 10.3102/016237370729970

Hembree, R. (1988). Correlates, causes, effects and treatment of test anxiety. *Review of Educational Research*, *58*(1), 47–77. doi: 10.2307/1170348

Hembree, R. (1990). The Nature, effects, and relief of mathematics anxiety. *Journal for Research in Mathematics Education*, *21*(1), 33-46. doi: 10.2307/749455

Higgins J.P.T., Deeks J.J., Altman D.G. (editors). Chapter 16: Special topics in statistics. In: Higgins J.P.T., Green S. (editors), *Cochrane Handbook for Systematic*

Reviews of Interventions Version 5.1.0 (updated March 2011). The Cochrane Collaboration, 2011. Available from www.handbook.cochrane.org.

*Hines, C. L. (2011). *The Effects of Expressive Writing on Anxiety, Mathematics Anxiety, Stress, Cognitive Processes and Psychological Processes on the Virginia Standards of Learning (SOL) on a Sample of Urban High School Students Failing Mathematics.* [Doctoral dissertation, Old Dominion University]

*Huang, X., & Mayer, R. E. (2016). Benefits of adding anxiety-reducing features to a computer-based multimedia lesson on statistics. *Computers in Human Behavior, 63,* 293-303. doi: 10.1016/j.chb.2016.05.034

Huntley, C. D., Young, B., Temple, J., Longworth, M., Smith, C. T., Jha, V., & Fisher, P. L. (2019). The efficacy of interventions for test-anxious university students: A meta-analysis of randomized controlled trials. *Journal of Anxiety Disorders*, *63*, 36-50. doi: 10.1016/j.janxdis.2019.01.007

*Husni, M. M. (2007). *Measuring the effect of anxiety reduction techniques on math anxiety levels in students enrolled in an HBCU college.* [Doctoral dissertation, University of Mississippi]

*Im, T. (2013). *The effects of emotional support and cognitive motivational messages on math anxiety, self-efficacy, and math problem solving.* [Doctoral dissertation, Florida State University]

*Insalaca, M. G. (2007). *The relationship between listening to music and high school students' levels of mathematics anxiety and mathematics achievement.* [Doctoral dissertation, Union Institute & University]

Ioannidis, J. P. A., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, *4*(3), 245-253. doi: 10.1177/1740774507079441

James, A. C., James, G., Cowdrey, F. A., Soler, A., & Choke, A. (2015). Cognitive behavioural therapy for anxiety disorders in children and adolescents. *The Cochrane database of systematic reviews*, *2015*(2), CD004690. doi: 10.1002/14651858.CD004690.pub4

Jameson, K. (2015). *A Brief DBT Treatment for Test Anxiety.* (Electronic Thesis or Dissertation). Retrieved from https://etd.ohiolink.edu/

*Jamieson, J. P., Peters, B. J., Greenwood, E. J., & Altose, A. J. (2016). Reappraising stress arousal improves performance and reduces evaluation anxiety in classroom exam situations. *Social Psychological and Personality Science, 7*(6), 579-587. doi: 10.1177/1948550616644656

Jaschik, S. (2020, June 15). *Research Universities Join the Test-Optional Movement*. Inside Higher Ed. https://www.insidehighered.com/admissions/article/2020/06/15/research-universities-join-test-optional-movement-least-year

*Kim, Y., Thayne, J., & Wei, Q. (2017). An embodied agent helps anxious students in mathematics learning. *Educational Technology Research & Development, 65*, 219-235. doi: 10.1007/s11423-016-9476-z

King, N., Mietz, A., Tinney, L., & Ollendick, T. (1995). Psychopathology and cognition in adolescents experiencing severe test anxiety. *Journal of Clinical Child Psychology, 24*(1), 49-54. doi:10.1207/s15374424jccp2401_6

Kirmayer, L. J., & Gómez-Carrillo, A. (2019). Culturally responsive clinical psychology and psychiatry: an ecosocial approach. *Cultural clinical psychology and PTSD, 2019*, 3-21.

Klausenitz, C., Hesse, T., Hacker, H., Hahnenkamp, K. & Usichenko, T. (2016). Auricular acupuncture for pre-exam anxiety in medical students: a prospective observational pilot investigation. *Acupuncture in Medicine, 34,* 90-94. doi: 10.1136/acupmed-2015-010887

Klein, A. (2015, November 30). ESEA reauthorization: The Every Student Succeeds Act explained. *Education Week*. Retrieved from http://blogs.edweek.org/edweek/campaign-k-12/2015/11/esea_reauthorization_the_every.html

Knappe, S., Beesdo-Baum, K., Fehm, L., Stein, M., Lieb, R., & Wittchen, H. (2011). Social fear and social phobia types among community youth: Differential clinical features and vulnerability factors. *Journal of Psychiatric Research, 45*(1), 111-120. doi:10.1016/j.jpsychires.2010.05.002

Leap, E. M. (2013). *A Quasi-Experimental Study Investigating the Effect of Scent on Students' Memory of Multiplication Facts and Math Anxiety.* [Doctoral dissertation, Robert Morris University]

LeBeau, R., Glenn, D., Liao, B., Wittchen, H., Beesdo-Baum, K., Ollendick, T., & Craske, M. (2010). Specific phobia: A review of the DSM-IV specific phobia and preliminary recommendations for DSM-V. *Depression and Anxiety*, *27*(2), 148–167. doi:10.1002/da.20655

Lobman, C. (2014). “I Feel Nervous... Very Nervous” Addressing test anxiety in inner city schools through play and performance. *Urban Education*, *49*(3), 329-359. doi: 10.1177/0042085913478621

McCoach, D. B. & Adelson, J. L. (2010). Dealing with dependence (Part I): Understanding the effects of clustered data. *Gifted Child Quarterly, 54*(2). 152-155. doi: 10.1177/0016986210363076

McDonald, A.S. (2001). The prevalence and effects of test anxiety in school children, *Educational Psychology, 21*(1), 89-101. doi: 10.1080/01443410020019867

Meijer, J. (2001). Learning Potential and Anxious Tendency: Test Anxiety as a Bias Factor in Educational Testing. *Anxiety, Stress & Coping*, *14*(3), 337-362. doi: 10.1080/10615800108248361

Mueller, C. M., & Dweck, C. S. (1998). Praise for intelligence can undermine children's motivation and performance. *Journal of Personality and Social Psychology*, *75*(1), 33. doi:10.1037/0022-3514.75.1.33

*Namwamba, J. O. (2013). *The effect of classical instrumental background music volume on performance in mathematics tests, self efficacy, and test anxiety of college students.* [Doctoral dissertation, Southern University and A&M College]

*Nelson, D. W., & Knight, A. E. (2010). The power of positive recollections: Reducing test anxiety and enhancing college student efficacy and performance. *Journal of Applied Social Psychology, 40*(3), 732-745. doi: 10.1111/j.1559-1816.2010.00595.x

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd edition). New York, NY: McGraw-Hill.

Osborne, J.W., Tillman, D., & Holland, A. (2009). Stereotype threat and anxiety for disadvantaged minorities and women. In J.C. Cassady (Ed.), *Anxiety in schools: The causes, consequences, and solutions for academic anxieties* (119-136). New York: Peter Lang Publishing.

*Park, D., Ramirez, G., & Beilock, S. L. (2014). The role of expressive writing in math anxiety. *Journal of Experimental Psychology: Applied, 20*(2), 103-111. doi: 10.1037/xap0000013

*Perez, A. I. (2005). *The impact of mathematics anxiety, gender, and mathematics achievement on ontogenetic indicators for Hispanic/Latino students in higher education mathematics classes.* [Doctoral dissertation, Texas A&M University]

Pigott, T. D., Valentine, J. C., Polanin, J. R., Williams, R. T., & Canada, D. D. (2013). Outcome-reporting bias in education research. *Educational Researcher*, *42*(8), 424-432. doi: 10.3102/0013189X13507104

Pinnock, G. (2014). *Intervention Program Data for Reducing Math Anxiety and Fostering Positive Social Change* (Doctoral dissertation). Retrieved from ProQuest Dissertations Publishing. (Accession No. 3646855)

Putwain, D. W. (2008). Test anxiety and GCSE performance: The effect of gender and socio-economic background. *Educational Psychology in Practice*, *24*, 319–334. https://doi.org/10.1080/02667360802488765

Putwain, D.W., & Best, N. (2011). Fear appeals in the primary classroom: Effects on test anxiety and test grade. *Learning and Individual Differences, 21,* 580-584. doi: 10.1016/j.lindif.2011.07.007

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. doi: 10.1080/03055698.2014.953914

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ran, L., Liu, X., Zhou, Y., Su, Y., Li, Y., & Wang, S. (2010). Effect of gum chewing on test anxiety. *Chinese Journal of Clinical Psychology, 18,* 731-734.

Reger, M.A., & Gahm, G.A. (2009) A meta-analysis of the effects of internet-and-computer-based cognitive-behavioral treatments for anxiety. *Journal of Clinical Psychology, 65,* 53-75. doi: 10.1002/jclp.20536

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

Seipp, B. (1991). Anxiety and academic performance: A meta-analysis of findings. *Anxiety research*, *4*, 27-41. doi: 10.1080/08917779108248762

Shadish, W. R., Hu, X., Glaser, R. R., Kownacki, R., & Wong, S. (1998). A method for exploring the effects of attrition in randomized experiments with dichotomous outcomes. *Psychological Methods, 3*(1), 3-22. https://doi.org/10.1037/1082-989X.3.1.3

Shadish, W.R., Robinson, L., & Lu, C. (1997). *ES: A computer program and manual for effect size calculation.* Memphis, TN: University of Memphis.

Sieber, J. E., O'Neil Jr., H. F. & Tobias, S. (1977). *Anxiety, Learning and Instruction*. Hillsdale, NJ: Erlbaum.

*Shen, E. (2009). *The effects of agent emotional support and cognitive motivational messages on math anxiety, learning, and motivation.* [Doctoral dissertation, Florida State University]

*Shobe, E., Brewin, A., & Carmack, S. (2005). A simple visualization exercise for reducing test anxiety and improving performance on difficult math tests. *Journal of Worry & Affective Experience, 1*(1), 34-52.

Sommer, M., & Arendasy, M. E. (2014). Comparing different explanations of the effect of test anxiety on respondents' test scores. *Intelligence*, *42*, 115-127. doi: 10.1016/j.intell.2013.11.003

Sommer, M., & Arendasy, M. E. (2015). Further evidence for the deficit account of the test anxiety–test performance relationship from a high-stakes admission testing setting. *Intelligence*, *53*, 72-80. doi: 10.1016/j.intell.2015.08.007

Spielberger, C. D. (1980). *Test Anxiety Inventory: Preliminary professional manual*. Menlo Park, CA: Mind Garden.

Spielberger, C. D., & Vagg, P. R. (1987). The treatment of test anxiety: A transactional process model. In H. M. van der Ploeg, R. Schwarzer, & C. D. Spielberger (Eds.), *Advances in test anxiety research (Vol*. *5*, p 179–186). Lisse, The Netherlands: Swets & Zeitlinger.

*Spielberger, S. L. (2015). *Effects of an expressive writing intervention aimed at improving academic performance by reducing test anxiety.* [Doctoral dissertation, Syracuse University]

Stöber, J. & Pekrun, R. (2004). Advances in test anxiety research, *Anxiety, Stress & Coping, 17*(3), 205-211. doi: 10.1080/10615800412331303225

Suinn, R., Edie, C., Nicoletti, J., & Sinelli, P. (1972). The MARS, a measure of mathematics anxiety: Psychometric data. *Journal of Clinical Psychology, 28*(3), 373-375.

Tempel, T. & Neumann, R. (2016). Taming test anxiety: The activation of failure-related concepts enhances cognitive test performance of test-anxious students. *The Journal of Experimental Education, 84*(4)702-722. doi: 10.1080/00220973.2015.1094649

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*(12), 2073-2083. doi: 10.1177/09567976221119391

*Thompson, H. R., Duvall, J., Padrez, R., Rosekrans, N., & Madsen, K. A. (2016). The impact of moderate-vigorous intensity physical education class immediately prior to standardized testing on student test-taking behaviors. *Mental Health and Physical Activity, 11*, 7-12. doi: 10.1016/j.mhpa.2016.06.002

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2020, October). *What Works Clearinghouse: Standards Handbook (Version 4.1)*. Retrieved from http://whatworks.ed.gov

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2020, October). *What Works Clearinghouse: Procedures Handbook (Version 4.1)*. Retrieved from http://whatworks.ed.gov

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2014, September). *Assessing Attrition Bias* [White paper]. https://ies.ed.gov/ncee/wwc/Document/243

Viechtbauer W (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1-48. doi: 10.18637/jss.v036.i03

Villar, J., Piaggio, G., Carroli, G., & Donner, A. (1997). Factors affecting the comparability of meta-analyses and largest trials results in perinatology. *Journal of Clinical Epidemiology*, *50*(9), 997-1002. doi: 10.1016/S0895-4356(97)00148-0

von der Embse, N., Barterian, J., & Segool, N. (2013). Test anxiety interventions for children and adolescents: A systematic review of treatment studies from 2000–2010. *Psychology in the Schools*, *50*(1), 57-71. doi: 10.1002/pits.21660

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. doi: 10.1016/j.jad.2017.11.048

Wallace, B.C., Small, K., Brodley, C.E., Lau, J. and Trikalinos, T.A. (2012). Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. In *Proc. of the ACM International Health Informatics Symposium (IHI)*, 819-824.

Wettlaufer, D.E. (2017). *The Effects of Massage on Student Stress* (Doctoral dissertation). Retrieved from ProQuest Dissertations Publishing. (Accession No. 10607659)

Wieland, L. R. (2011). *Brainology and math achievement.* [Doctoral dissertation, Fairleigh Dickinson University]

Wisinger, S. (2010). *The effects of two anxiety reducing interventions on Algebra I test scores for a sample of rural high school students.* [Doctoral dissertation, Old Dominion University]

Yahav, R. & Cohen, M. (2008). Evaluation of a cognitive-behavioral intervention for adolescents. *International Journal of Stress Management. 15,* 173-188. doi: 10.1037/1072-5245.15.2.173

Yeo, L.S., Goh, V.G. & Liem, G.A.D. (2016). School-based intervention for test anxiety. *Child & Youth Care Forum, 45*, 1-17. doi: 10.1007/s10566-015-9314-1

Yerkes, T.M. & Dodson, J.D. (1908). The relation of strength stimulus to rapidity of habit formation. *Journal of comparative neurology and psychology, 18(5), 459-482.* doi: 10.1002/cne.920180503

Young, S., Premji, Z. & Engelbert, M. Unit 3: Searching the Literature. In J. C. Valentine, J. H. Littell, & S. Young (Eds.), *Systematic reviews and meta-analysis: A Campbell Collaboration online course*. Open Learning Initiative, 2021. Available from https://oli.cmu.edu/courses/systematic-reviews-and-meta-analysis/

Zeidner, M. (1998). *Test anxiety: The state of the art*. New York: Plenum.

Zeidner, M., & Mathews, G. (2005). Evaluation anxiety: Current theory and research. In A. Elliot and C. Dweck (Eds.), *Handbook of competence and motivation*, *141-163*. New York, NY: Guilford Publications.

Zeidner, M., (2014). Test Anxiety. In P. Emmelkamp and T. Ehring (Eds.), *The Wiley Handbook of Anxiety Disorders, 581-595*. West Sussex, UK: John Wiley & Sons, Ltd. doi: 10.1002/9781118775349.ch28

Zenner, C., Herrnleben-Kurz, S., & Walach, H. (2014). Mindfulness-based interventions in schools—a systematic review and meta-analysis. *Frontiers in Psychology, 5, 1-20*. doi: 10.3389/fpsyg.2014.00603

Zlomke, J. M. (2007). *Test anxiety in nursing students* (Doctoral dissertation). Retrieved from ProQuest Dissertations Publishing. (Accession No. 1443295)

Zohar, D. (1998). An additive model of test anxiety: Role of exam-specific expectations. *Journal of Educational Psychology, 90*, 330-340. doi:10.1037/0022-0663.90.2.330

APPENDIX A

Code and Data

https://github.com/TJReece/Meta-Analysis-of-Therapeutic-Interventions-for-the-Treatment-of-Test-Anxiety.git

CURRICULUM VITA

Thomas John Reece, M.A.

Work: Jefferson County Public Schools
VanHoose Education Center
3332 Newburg Road
Louisville, KY 40218
Phone: (502) 485-3276
Thomas.Reece@jefferson.kyschools.us

Education

2015-2023 **University of Louisville**, Louisville, KY.

Doctor of Philosophy in Counseling and Personnel Services

Specialization in Educational Psychology, Measurement, and Evaluation

Doctoral Candidate

2006-2009 **Western Kentucky University**, Bowling Green, KY.

Masters of Arts in Psychology (Clinical Concentration)

Magna Cum Laude

2004-2006 **Western Kentucky University**, Bowling Green, KY.

Bachelor of Arts in Psychology

Magna Cum Laude

Work Experience:

2019-Present **Data & Research Coordinator, Jefferson County Public Schools**

Supervisors: Bo Yan, Ph.D.

Conducts data analysis, program evaluations, and provides technical assistance to support district personnel, district research, district systems, program evaluations, and schools. Data analysis primarily consists of statistical analysis of internal files. The analyses are used as one of the components in the district decision-making process. Supports district-level efforts related to the state accountability system, including quality control, communication of accountability data to school- and district-level stakeholders, and production of data discrepancy reports. Produces and maintains a variety of automated reports directed towards a variety of stakeholders. Conduct program evaluations.

2017-2018 **Data Analyst, Jefferson County Public Schools**

Supervisor: Joe Prather, Ed.D.

Conducts data analysis and provides technical assistance to support district personnel, district research, district systems, program evaluations, and schools. Data analysis primarily consists of statistical analysis of internal files. The analyses are used as one of the components in the district decision-making process.

2017 **Data Management and Research Intern, Jefferson County Public Schools**

Supervisors: Joe Prather, Ed.D. & Florence Chang, Ph.D.

Involved in research supporting the data analysis and evaluation needs of the district. Related tasks include: psychometric analysis of district-wide surveys, evaluation of program effectiveness, fulfilling internal and external data and analysis requests.

Consulting Experience:

2022-2023 **UnboundEd Learning Inc.**

Analyze internal quantitative and qualitative data.

2018-Present **Development Services Group**

Conduct reviews of quantitative research studies according to What Works Clearinghouse study review standards.

Research Experience:

2015-2018 **What Works Clearinghouse**

Research Director: Jeff Valentine, Ph.D.

Involved in evaluating research in postsecondary education. Responsibilities include screening abstracts for meta-analyses, working on projects to improve WWC quality metrics, and reviewing studies.

2010-2015 **Clinical and Applied Research Group, Western Kentucky University**

Research Director: Rick Grieve, Ph.D.

Supervise undergraduate and graduate research primarily focusing on sports fan identification, and male body image. Responsibilities include aiding in the construction of research projects, conducting statistical analysis, and coordinating online resources for the research group.

2006-2009 **Office of Applied Research and Analysis, Western Kentucky University**

Research Advisor: John Bruni, Ph.D.

Involved in institutional research, database construction, and data analysis. Projects included: analyzing retention rates of African American males, institutional assessment of adult learning goals, and assisting faculty in the development and analysis of research projects.

2005-2009 **Clinical and Applied Research Group, Western Kentucky University**

Research Advisor: Rick Grieve, Ph.D.

Involved in research on body image and muscle dysmorphia.

Peer-reviewed Publications:

Roelfs, D. J., Högnäs, R. S., Shor, R., Moore, C., & Reece, T. (2017) J-Curve? A Meta-Analysis and Meta-Regression of Parity and Parental Mortality. *Population Research and Policy Review.* doi: 10.1007/s11113-016-9421-1

Grieve, F. G., Jackson, L., Reece, T., Marklin, L., & Delaney, A. (2008). Correlates of social physique anxiety in men. *Journal of Sport Behavior, 31*, 329-337.

Book Chapters:

Lee, B.N., Reece, T.J., & Grieve, F.G. (2017). Body dysmorphic disorder and gender. In K.L. Nadal (Ed.), *The SAGE Encyclopedia of Psychology and Gender* (pp. 216-219). Thousand Oaks, CA: SAGE Publications. doi: 10.4135/9781483384269.n

Presentations at Professional Meetings:

Invited Talks:

Reece, T. (February, 2017). *IRT Basics.* Presentation for the Data Management, Planning, and Program Evaluation department of Jefferson County Public Schools, Louisville, Ky.

Cuisinier, E., Bundy, M., Reece, T., Schulenberg, S. (September, 2015). S*trategies for teaching abroad.* Panel discussion at the annual faculty meeting of the Kentucky Institute for International Studies, Bowling Green, KY.

Paper Presentations:

Immekus, J. C., Lau, T., & Reece, T. (2017). *An investigation of the psychometric properties of an observational checklist of children's emergent literacy skills within a multilevel item response theory framework*. Paper presented at the annual conference of the American Education Research Association at San Antonio, TX.

Reece, T., Crawford, B.F., & Immekus, J.C. (March, 2016). *Jumping to Conclusions about Item Bias: Analysis of Differential Item Functioning in the Jumpstart School Success Checklist.* Presented at the annual Spring Research Conference at the University of Kentucky.

Eovino, J., Grieve, F.G., Reece, T., Morris, B., King, S., Kirsch Hiltz White, C., West, H., Acree, J., Simpson, S., & Gooch, A. (2014, October). *And the horse you rode in on too!: Sport fan perspectives of collective punishment.* Presented at the annual conference of the Academy of Business Research, Las Vegas, NV.

Reece, T.J. (2008, April). *Culture shock in international students.* Presented at the annual Student Research Conference at Western Kentucky University, Bowling Green, KY.

Poster Presentations:

Reece, T. J., & Bundy, M., Asher, J. (2015, November). *"The hardest place to be" - Psychology student reactions to visiting a concentration camp.* Poster presented at the annual conference of the National Collegiate Honors Council at Chicago, Illinois.

Reece, T. J., & Bundy, M. (2015, January). *"The hardest place to be" - Psychology student reactions to visiting a concentration camp.* Poster presented at the annual conference of the National Institute for the Teaching of Psychology at St. Pete Beach, Florida.

Reece, T. J. (2014, August). *Religiosity's relationship with homosexuality: It's not as straight as we thought.* Poster presented at the annual conference of the American Psychological Association at Washington D.C.

Reece, T. J., West, H., Grieve, F., Aldridge, L., Young, A., Armstrong, M., Isbill, A., Kirsch-Hiltz White, C., McCarthy, B., White, M., Eovino, J., & Cyr, C. (2013, May). *Man's search for body satisfaction.* Poster presented at the annual conference of the Association for Psychological Science at Washington D.C.

Reece, T. J., & Grieve, F. (2012, May). *PTSD and malingering in a Bosnian refugee sample.* Poster presented at the annual conference of the Association for Psychological Science at Chicago, IL.

Reece, T. J. (2011, May). *A person-centered exploration of religious dimensions.* Poster presented at the annual conference of the Association for Psychological Science at Washington D.C.

Reece, T. J. (2010, May). *Three statistical personality types.* Poster presented at the annual conference of the Association for Psychological Science at Boston, MA.

Reece, T. J. (2009, March). *A more nuanced look at religious orientation and homonegativity.* Poster presented at the annual meeting of the Midwestern Psychological Association, Chicago, IL.

Reece, T. J. (2009, February). *A more nuanced look at religious orientation and homonegativity.* Poster presented at the annual Student Research Conference at Western Kentucky University, Bowling Green, KY.

Reece, T. J., Pegg, P. (2008, April). *The structure of fears.* Poster presented at the annual meeting of the Middle Tennessee Psychological Association, Clarksville,TN.

Reece, T. J. (2005, April). *Development of a measure of religious faith and its relationship with nonmarital sexual acceptance.* Poster presented at the annual meeting of the Middle Tennessee Psychological Association, Murfreesboro, TN.

Teaching Experience:

Fall 2008 - Present **Adjunct Instructor**, *Western Kentucky University*

2008 – Present **Psy100: Introduction to Psychology:** *Course Instructor*

Freshman-level undergraduate Introduction to Psychology course; developed and carried out curriculum including lectures, class discussions, group projects, individual student papers, and multiple-choice and essay exams. Taught as a face-to-face class and as a web course.

2009 – 2020 **Psy199: Developmental Psychology:** *Course Instructor*

Freshman-level undergraduate course on lifespan development; developed and carried out curriculum including lectures, class discussion, individual student projects/papers, and multiple-choice exams. Taught as a face-to-face class and as a web course.

2013 – 2014 **Psy350: Social Psychology:** *Course Instructor*

Junior-level undergraduate social psychology course; developed and carried out curriculum including discussion board, individual student projects/papers, and multiple-choice quizzes. Taught as a web course.

Summer 2014 **Psy299: The Social Psychology of Evil – Psychology and The Holocaust:** *Course Instructor*

Special topics course focused on the social psychological examination of the Holocaust and other genocides. Developed and carried out curriculum including lectures, class discussion, individual student projects/papers, and group presentations. Taught with the Kentucky Institute for International Studies in Bregenz, Austria.

Summer 2016 **Psy299: The Psychology of Good and Evil:** *Course Instructor*

Special topics course focused on the psychological examination of good and evil. Developed and carried out curriculum including lectures, class discussion, individual student projects/papers, and group presentations. Taught with the Kentucky Institute for International Studies in Bregenz, Austria.

2014 – 2015 **Psy455: Introduction to Clinical Psychology:** *Course Instructor*

Senior-level undergraduate course in clinical and applied psychology; developed and carried out curriculum including discussion board, individual student projects/papers, and multiple-choice quizzes. Taught as a web course.

Spring 2010 - 2021 **Part-time Instructor**, *Southcentral Kentucky Community and Technical College*

2010 - 2021 **Psy110: General Psychology:** *Course Instructor*

Freshman-level undergraduate Introduction to Psychology course; developed and carried out curriculum including lectures, class discussions, group projects, individual student papers, and multiple-choice and essay exams. Taught as a face-to-face class and as a web course.

2010 - 2021 **Psy223: Developmental Psychology:** *Course Instructor*

Sophomore-level undergraduate course on lifespan development; developed and carried out curriculum including lectures, class discussion, individual student projects/papers, and multiple-choice and essay exams, and individual student papers. Taught as a face-to-face class and as a web course.

Spring 2015 - 2021 **Psy298: Essentials of Abnormal Psychology:** *Course Instructor*

Sophomore-level undergraduate course on abnormal psychology; developed and carried out curriculum including lectures, class discussion, individual student projects/papers, and multiple-choice and essay exams, and individual student papers. Taught as a web course.

Fall 2014 - Spring 2015 **Facilitator/Instructor**, *Eastern Kentucky University*

2014 **Psy377: Psychology of Adoption:** *Course Facilitator*

Junior-level undergraduate course on the developmental, familial, cultural, and psychological factors in the adoption experience; participated in and graded discussion board activities and graded weekly short answer essay quizzes. Taught as a web course.

2015 **Psy308: Abnormal Psychology:** *Course Facilitator*

Junior-level undergraduate course on abnormal psychology; participated in and graded discussion board activities, and graded weekly short answer essay quizzes.

2016 **Psy408: Child Psychopathology:** *Course Instructor*

Senior-level undergraduate course on abnormal psychology, with a special focus on problems experienced during childhood; carried out curriculum including lectures, class discussion, individual student projects/papers, and multiple-choice and essay exams, and individual student papers. Taught as a web course.

University Service:

University Service:

2013-2018 Applied Statistics Center Faculty Mentor: *Western Kentucky University*

Duties include advising and mentoring student consultants working for the Applied Statistics Center. Under the guidance of affiliate faculty members, students in the Applied Statistics Center collaborate with academic researchers to find statistical solutions to problems.

College/Departmental Service:

2011-2015 Freshman/Sophomore Advisor: *Western Kentucky University*

Duties include assisting students with the selection of classes and mentoring students through their first two years of college. A special focus on improving student retention. Additional duties include attending incoming freshman orientation (Academic Transitions Program), and training and serving as a resource for junior advisors.

Clinical Experience:

2007 - 2008 **Clinical Internship,** Western Kentucky University

My clinical internship consisted of working at Western Kentucky University's Psychology Training Clinic and the Office of International Programs. My duties at both internship sites included development and implementation of treatment plans, individual and group supervision, staff meetings, and attendance to didactic trainings and group presentations.

Psychology Training Clinic

Supervisor: Rick Grieve, Ph.D.

I conducted psychological assessments primarily to assist with the diagnosis of academic difficulties in a college student population.

Office of International Programs

Supervisors: Phil Pegg, Ph.D., Robin Borczon

Responsible for providing individual psychotherapy to international students studying at Western Kentucky University suffering from a variety of psychological disorders. My responsibilities included treatment planning and implementation, and consultation with university staff and support personnel. I was also responsible for weekly psychoeducational group therapy with newly arrived international students.

2007 **Psychology Training Clinic**, Western Kentucky University
Supervisor: Melissa Hakman, Ph.D.
Conducted individual psychotherapy with college students. Treatment focused on using cognitive-behavioral techniques targeting behavioral disturbances.

2006 - 2008 **Psychology Training Clinic,** Western Kentucky University
Supervisors: Melissa Hakman, Ph.D.
Rick Grieve, Ph.D.
Conducted cognitive and personality assessments with children, adolescents, and college-aged students.

2005 **Rivendell Behavioral Health Services**, Bowling Green, Kentucky
Supervisor: Jennifer A. Miller, LCSW
Primary experience in therapy for substance dependence and abuse, general record keeping, and administrative and treatment team decision making.

Certifications:

2018; What Works Clearinghouse Group Design Standards Certified Reviewer (Version 4.1)

2017; What Works Clearinghouse Group Design Standards Certified Reviewer (Version 4.0)

Positions and Honors:

2012; Awarded Master Advisor Certificate, Western Kentucky University

2010; Inducted into the Honorable Order of Kentucky Colonels

2009; Graduated Magna Cum Laude, Western Kentucky University

2006; Graduated Magna Cum Laude, Western Kentucky University

2006; Psi Chi National Honor Society (Treasurer)

Professional Memberships:

2015 – 2020, American Educational Research Association

2010 – 2017, Association for Psychological Science

2010 – Present, Society for the Teaching of Psychology

2006 – Present, Psi Chi National Honor Society

2005 – 2017, American Psychological Association