Journal of Anxiety Disorders 63 (2019) 36–50

Contents lists available at ScienceDirect

Journal of Anxiety Disorders

journal homepage: www.elsevier.com/locate/janxdis

Review

The efficacy of interventions for test-anxious university students: A meta-analysis of randomized controlled trials

Christopher D. Huntley[a,*], Bridget Young[a], James Temple[a], Melissa Longworth[b], Catrin Tudur Smith[a], Vikram Jha[c], Peter L. Fisher[a]

[a] University of Liverpool, UK
[b] Lancaster University, UK
[c] Apollo Hospitals Education and Research Foundation, Chennai, India

ARTICLE INFO

Keywords:
Test anxiety
Meta-analysis
Academic performance
Interventions

ABSTRACT

Test anxiety (TA) is highly distressing and can significantly undermine academic performance. Many randomized controlled trials (RCTs) of interventions for university students with TA have been conducted, but there has been no systematic review of their efficacy. This meta-analysis examines the efficacy of interventions for test-anxious university students in: (i) reducing TA, and (ii) improving academic performance. We searched for RCTs published in English language peer-reviewed journals. Forty-four RCTs met our eligibility criteria (n = 2,209). Interventions were superior to control conditions at post-treatment for reducing TA (g = −0.76) and improving academic performance (g = 0.37). Interventions were superior to control conditions at follow-up. Subgroups analyses found most support for behaviour therapy. Cognitive-behavioural therapy, study skills training, and combined psychological and study skills training interventions show promise but lack evidence for their longer-term efficacy, and results are based upon a small number of studies. Evidence of publication bias was found and poor quality of reporting meant that confidence in results should be moderated. Future RCTs should be conducted and reported with greater rigour, have larger samples, and examine newer interventions.

1. Introduction

Examinations are a prime concern for university undergraduate students (Knappe, Beesdo-Baum, Fehm, Stein, Lieb, & Wittchen, 2011) with around 20–25% of undergraduate university students estimated to be highly test-anxious (Hahne, Lohmann, & Krzyszycha, 1999, cf. Hill & Wigfield, 1984; Naveh-Benjamin, Lavi, McKeachie, & Lin, 1997; Neuderth, Jabs, & Schmidtke, 2009; Saravanan, Kingston, & Gin, 2014). Anxiety about examinations – or *test anxiety* (TA) – concerns the "phenomenological, affective, and behavioural responses that accompany concern about the possible negative consequences of poor performance in an examination or other performance-evaluative situations" (Zeidner & Matthews, 2005, p. 142). TA is a situation-specific anxiety disorder that consists of two main dimensions: a cognitive dimension, typically labelled 'Worry', that consists of perseverative thinking about the consequences of failure and test-irrelevant thinking, and an affective dimension labelled 'Emotionality' that refers to physiological arousal (e.g., muscle tension, sweating, and heart rate accelerations) in test situations.

Meta-analyses conclude TA is associated with poorer examination and academic performance. In the first meta-analysis Hembree (1988) analyzed 562 studies published between 1952 and 1986. Data were synthesized across age groups and educational stages, i.e., from Grade 1 (6–7 years) up to university students (18 years and older). TA and academic performance were significantly negatively correlated (r = −0.29). Worry was more strongly correlated with poor performance (r = −0.31) than Emotionality (r = −0.15). A subsequent meta-analysis (Seipp, 1991) analyzed data from 126 studies published between 1975 and 1988, with data were synthesized across age groups and educational stages. TA and academic performance were significantly negatively correlated (r = −0.23). Again, worry was more strongly associated with poor performance (r = −0.22) than Emotionality (r = −0.15); however, analyses revealed significant heterogeneity within Worry but not Emotionality suggesting other person-situation variables influence the relationship between Worry and academic performance (e.g., the test 'stakes', such as whether it is a formative or summative examination, might influence the degree of worry experienced for students). The most recent meta-analysis (von der Embse, Jester, Roy, & Post, 2018) analyzed data from 238 studies published between 1988 and 2018. The main results were aggregated

* Corresponding author at: School of Medicine, University of Liverpool, Brownlow Hill, Liverpool, L69 3GB, UK.
E-mail address: C.Huntley@liverpool.ac.uk (C.D. Huntley).

https://doi.org/10.1016/j.janxdis.2019.01.007
Received 17 May 2018; Received in revised form 27 November 2018; Accepted 28 January 2019
Available online 06 February 2019
0887-6185/

across age and TA was negatively correlated with academic performance. Worry was more strongly associated with TA than emotionality. In the sub-analysis of university students (53 studies; $n = 20{,}849$) higher levels of TA were associated with poorer academic performance ($r = -0.27$). In summary, results from the three meta-analyses support the contention that TA adversely impacts academic performance.

TA adversely affects student mental health. High-test-anxious students report poorer mental health (Depreeuw & De-Neve, 1992) and are more likely to dropout or repeat a year of study (Schaefer, Mattheß, Pfitzer, & Köhle, 2007; cf. Neuderth, Jabs, & Schmidtke, 2009) compared to low-test-anxious students. TA is highly comorbid with other mental health disorders, particularly depression and social anxiety (Herzer, Wendt, Hamn, 2014; Kavakci, Guler, & Cetinkaya, 2014). Furthermore, in a national study of suicides in young people in England between January 2014 and April 2015, academic pressures were reported to be a potentially important factor in suicides amongst those in education, with 29% of suicide cases involving students facing examinations or examination results at the time of death (Suicide in Children and Young People, 2016). TA is frequently cited as one the principal reasons for students seeking access to mental health and university student support services (Rückert, 2015).

Psychological interventions have been most commonly applied to treating TA in undergraduate students (Zeidner, 1998). TA interferes with studying and test taking and therefore interventions which alleviate symptoms of TA should improve performance. Psychological interventions for TA fall into two broad categories: behaviour therapy (BT) and cognitive-behavioural therapy (CBT). BT focuses on reducing the affective or emotionality dimension of TA, typically through relaxation techniques. The main interventions were progressive muscle relaxation (Jacobson, 1938) which involves the sequentially relaxation muscle groups or systematic desensitization (Wolpe, 1958) which involves counterconditioning techniques to teach muscle relaxation while visualizing a hierarchy of increasingly stressful TA-related scenes. Further relaxation methods tested were cue-controlled relaxation (Russell, Wise, & Stratoudakis, 1976) involving training to relax muscle groups in response to self-induced cues, or exposure-based interventions such as implosive therapy or flooding (Wolpe, 1969) that involve real or imagined exposure to the feared object or situation and are underpinned by classical conditioning principles. CBT-based interventions aim to modify cognitions and behaviours. The most prominent CBT intervention for anxiety disorders is based upon the work of Beck and colleagues (Beck, Emery, & Greenberg, 1986). Here, the aim is to identify negative automatic thoughts (NATs) and cognitive distortions which lead to and maintain anxiety. An example of a cognitive distortion is 'generalisation', whereby individuals draw broad conclusions based on a single event (e.g., a student getting a low grade and thinking they are completely inadequate). Therapists guide clients to recognize and challenge their NATs by first labelling the cognitive distortion contained within NAT and then followed by logical disputation e.g., exploring the evidence for and against the NAT.

Another class of interventions applied to treating TA are study skill training (SST) interventions. These are based upon the assumption that inadequate test preparation and test taking skills impair academic performance and that TA is an epiphenomenon arising from an individual's appraisal that they are inadequately prepared for the test. SST interventions therefore seek to enhance the learning and test-taking abilities to increase confidence and reduce anxiety in test taking situations so that individuals feel more confident about meeting their performance goals (Zeidner, 1998). SST interventions most commonly consist of two components (e.g. Dendato & Diener, 1986), one focusing on effective ways of learning and encoding study material (e.g. deeper-levels of learning versus rote-learning), with the other component focusing on effective strategies during examinations (e.g. allocating more time to those questions which represent a greater proportion of the total examination score). Intervention packages have also been designed that combine psychological (BT, CBT) and SST interventions in an attempt to address the psychological causes of TA and improve effective study and test-taking skills (e.g., McCordick, Kaplan, Smith, and Finn, 1981).

Finally, pharmacological approaches target the neurochemical origins of anxiety. Medications used to treat TA includes tranquilizers, beta-blockers, anti-depressants, Selective Serotonin Reuptake Inhibitors (SSRIs), Serotonin-Norepinephrine Reuptake Inhibitors (SNRIs), and benzodiazepines, though there is scant research on their application and efficacy (e.g., Brewer, 1971).

1.1. Previous meta-analytic reviews of interventions for TA

Two meta-analyses have examined the efficacy of interventions for TA (Ergene, 2003; Hembree, 1988). Based on analyses of 137 studies ($n = 7641$ students) published between 1950 and 1986, Hembree (1988) concluded that BT and CBT were effective in reducing TA and improving academic performance. Ergene (2003) examined the efficacy of interventions in 56 studies ($n = 2428$ students) published between 1950 and 1998 focusing only on controlled trials, and concluded that BT, CT, and Combined psychological and SST interventions were effective in reducing TA (academic performance was not examined). Neither review included any pharmacological studies.

However, both meta-analyses were not conducted with the rigour expected of current systematic reviews, casting doubt on their conclusions. Both included and synthesized results from studies on primary/elementary school children through to university undergraduates. Clearly, the presentation of symptoms will vary considerably across age ranges, and therefore the suitability and efficacy of interventions may also vary considerably. Though Ergene (2003) reported a summary effect size of interventions for test-anxious undergraduates ($d = -0.68$, 95% CI -0.77 to -0.59), no summary statistics were reported by Hembree (1988), and neither review examined the efficacy of different treatment approaches for undergraduate students. Therefore, the efficacy of specific interventions for specific patient groups is unknown. It cannot be assumed that an efficacious intervention in one patient group (e.g., children under 10 years) is equally efficacious in another group (e.g., university students over 18 years), and given that analyses in both reviews synthesized across age groups it is possible that an efficacious interventions for one group may mask or moderate the effect for another group.

Another limitation is that both reviews used a fixed-effect model, which assumes that all studies are estimating a single effect and that any variation between studies is a result of sampling error. In effect, the use of a fixed-effect model implies that researchers saw the studies as functionally identical; but this cannot be the case given the wide age differences of samples across included trials. Treatment trials of TA have frequently included more than two treatments arms (e.g. Meichenbaum, 1972) but neither review reports how these trials are dealt with within their analyses, and therefore some participants, such as those in control conditions, may have been included more than once within their analyses. This can bias results, by selectively increasing sample size for those studies with more than two conditions, giving them greater weight within data syntheses, and reducing confidence intervals around the effect size estimates (Borenstein, Hedges, Higgins, & Rothstein, 2009).

The final major limitation of these two reviews (Ergene, 2003; Hembree, 1988) concerns the poor quality of reporting when viewed through the lens of modern reporting standards i.e., the Preferred Reporting Standards for Systematic Reviews and Meta-analyses (PRISMA; Moher, Liberati, Tetzlaff, & Altman, 2009). Neither review identifies to the reader which trials were included and their basic characteristics (e.g., interventions delivered, delivery format, number of sessions), and nor are forest plots presented to convey the variation in outcomes between studies and which studies contributed most weight to the summary effect size. Additionally, a priori protocols are not available for either review.

Given these serious problems with the methodological and reporting

standards of the two reviews above, the efficacy of interventions for test-anxious university students is unknown. Therefore, a new up-to-date review of the efficacy of interventions focusing specifically on undergraduate students is required.

1.2. Aims of this review

This review has three main aims, to examine: (1) the efficacy of interventions for reducing TA in university undergraduate students (henceforth 'students'), (2) potential moderators of treatment effect (see Section 1.3 for more detail) and, (3) the efficacy of interventions for improving academic performance.

1.3. Moderators of intervention effect

We planned to examine the following sample-dependent moderators of intervention effect: gender and pre-treatment TA severity. Given females report greater TA severity than males (Cassady & Johnson, 2002; Zeidner, 1998), we hypothesized that studies with higher proportions of female participants will have poorer treatment outcomes. We also hypothesized that pre-treatment severity would be associated with poorer treatment outcomes.

Several methodological moderators of intervention effect were also examined: mode of treatment delivery (i.e., group vs. individual), treatment dosage (i.e., number of hours of treatment), and manualization of interventions (manualized vs. non-manualized). We hypothesized that group interventions would be associated with better treatment outcomes than individual interventions, in-line with Ergene (2003) findings that group interventions produced a larger effect size than individual interventions. Ergene (2003) also examined the effect of treatment hours on effect size but no clear dose-response pattern emerged, and we therefore hypothesized no significant relationship. Manualization may improve the quality of treatment delivery (Addis, Cardemil, Duncan, & Miller, 2006), and therefore it was hypothesized that manualized interventions would be associated with better treatment outcomes.

2. Methods

The review was conducted and reported in accordance to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA; Moher et al., 2009) guidelines. Development of the study protocol (Huntley, Young, Jha, & Fisher, 2016) was guided by Cochrane protocol review guidelines (Higgins & Green, 2011) and was registered with PROSPERO (CRD42016035859).

2.1. Identification and selection of studies

2.1.1. Eligibility criteria

Included studies: (i) included only test-anxious university undergraduate students (self-selected and/or meeting criteria for severity of TA); (ii) examined psychological, SST, pharmacological interventions, and/or combined intervention packages; (iii) had at least one comparator that was a control condition (no treatment, waitlist, or psychological/pill placebo); (iv) reported a primary outcome of TA severity on a valid measure; (v) used random assignment of participants to conditions; and (vi) was published in a peer-reviewed journal written in English.

2.1.2. Identification of studies

Six electronic databases – Cochrane Central Register of Controlled Trials (CENTRAL), Educational Resources Information Center (ERIC), MEDLINE, PsycINFO, Scopus, and Web of Science – were searched using variations of the relevant search terms covering: TA; undergraduate student; psychological, SST, or pharmacological interventions; randomized controlled trial. An example search protocol for the Scopus database was previously presented (Huntley et al., 2016) and the final search protocol can be located at PROSPERO. The first search period covered all years up to April 2016. A further search was conducted, covering the period April 2016 up to May 2017.

A manual search was conducted, checking reference lists of previous meta-analytic reviews and primary articles. CDH conducted all searches. Articles identified through the literature searches were then combined by CDH into a single database and duplicates removed.

2.1.3. Screening of studies

Two researchers (JT and ML) independently screened studies against our eligibility criteria, with discrepancies resolved by CDH. First, studies were screened by title and abstract. Full-text copies of the remaining included studies were then retrieved and screened using tool with items for each of our eligibility criteria. Studies could be then marked as 'Included', 'Maybe', or 'Excluded'.

2.2. Data extraction and risk of bias assessment

Predefined information was independently extracted by JT and ML from the study reports using a modified Cochrane data extraction form. Risk of bias (RoB) was assessed by Cochrane's RoB tool (Higgins et al., 2011) that assesses the following domains: random sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessment, incomplete outcome data, and selective reporting. Each domain is scored either: 'low risk of bias', 'unclear', or 'high risk of bias'. If relevant information was not reported, study authors were contacted by CDH and the missing study information requested. If there was no response within two weeks, a reminder email was sent. Following this, no further efforts were made to obtain information.

2.3. Planned analyses

Meta-analyses were conducted using Revman 5.3.5 (Cochrane Collaboration, 2014, Review Manger, 2014), while tests of publication bias and meta-regression analyses were conducted in R 3.3.3 (R Core Team, 2013) using the 'metafor' package (Viechtbauer, 2010).

2.3.1. Data synthesis

A random-effects model was used. TA severity is measured on a continuous scale, but a wide variety of self-report outcome measures have been used across studies. Therefore, between-study standardized mean differences (SMD) with 95% confidence intervals were calculated for each pairwise comparison (active intervention versus control group) and for the pooled effect, using Cohen's effect size with Hedges' correction for small sample bias i.e., Hedges' g (Hedges, 1981).

Studies with more than two treatment comparisons were included. However, independence between multiple comparisons within the same study cannot be assumed (Borenstein, Hedges, Higgins, & Rothstein, 2009). To adjust for the potential bias the control group n is divided by the number of active comparator conditions (Higgins & Green, 2011) ensuring each participant is only included once in the analyses. Data from the primary TA outcome measure reported in each study were used (if no primary outcome stated, the first reported TA outcome was then used).

The magnitude of heterogeneity was assessed using the I^2 statistic, with values greater than 50% indicative of at least moderate heterogeneity (Higgins, Thompson, Deeks, & Altman, 2003). Subgroups analyses were used to examine the relative efficacy of different interventions at post-treatment and follow-up. Using past reviews as a guide, the treatment categories were decided as the following: BT, CBT, SST, and Combined. Combined interventions consisted of both psychological (i.e., BT or CBT) and SST components.

The same procedure outlined above was conducted for the academic performance. Academic performance indices used here included either

post-treatment examination results or grade point average (GPA). GPA is summary statistic representing the average value of accumulated grades earned across course modules. Most commonly, GPA is based on a 0 to 4.0 scale (A = 4.0, B = 3.0, C = 2.0, D = 1.0, Fail = 0.0) where a 4.0 GPA average represents the highest score.

Outliers were identified via visual inspection of the forest plot (i.e., where study effect estimates, and their 95% confidence interval, had little overlap with other study effect size estimates). Outliers were then removed and results compared with original estimates.

2.3.2. Moderator analyses

Moderators that we wish to investigate have been discussed previously (see Section 1.3). A minimum of 10 studies for each study-level variable is recommended (Higgins & Green, 2011). Therefore, if heterogeneity was present within any of the subgroups, and there were sufficient studies, moderators of treatment outcome were examined via mixed-effects meta-regression (i.e. studies are treated as being a random sample but the moderators are treated as being fixed-effect). Regarding manualization, a study had to explicitly cite a manual or provide a detailed description within the manuscript to be coded as "manualized"; else they were coded as non-manualized.

2.3.3. Sensitivity analyses

Sensitivity analyses were planned to assess the robustness of results, with RCTs considered at high risk of bias removed and then results compared to original analyses.

2.3.4. Assessment of publication bias

Publication bias was assessed using Egger's test (Egger, Smith, Schneider, & Minder, 1997) and funnel plots of the effect sizes against the standard error (SE).

2.4. Quality of evidence

Overall quality of evidence was assessed with Grading of Recommendations Assessment, Development and Evaluation (GRADE; Guyatt et al., 2008) using their GRADEpro Guideline Development Tool (GDT). Five domains of quality are assessed: risk of bias, inconsistency of trial results, indirectness of measurement, imprecision of effect size estimates, and the likely impact of publication bias. Evidence quality is graded on a 4-point scale from 'very low quality' (i.e., limited confidence in effect size estimate) to 'high quality' (i.e., high confidence in the effect size estimate).

3. Results

3.1. Identification and selection of studies

Fig. 1 shows the PRISMA flowchart. Our search identified 1120 papers. After removing 253 duplicate records, 867 records were independently screened (JT, ML) by title and abstract, with 692 records excluded. Full-text copies of the remaining 175 papers were obtained, with multiple papers corresponding to a single study grouped together, resulting in 168 unique studies. Full-texts were independently screened (JT, ML), with 78 studies meeting the inclusion criteria; inter-rater reliability was good (Kappa = 0.60). The list of studies excluded (including reasons for exclusion) at the full-text screening stage can be found in Appendix A (online only). The 78 studies were then examined to check if data necessary for inclusion in the meta-analyses was reported. Thirty-four studies did not report sufficient information; the study authors were contacted and asked to supply the required information. Although 16 (47%) of the study authors responded, none could provide the necessary information; primarily because their studies were conducted more than 20 years ago and data was no longer accessible. Forty-four studies were therefore included in the meta-analysis.

3.2. Characteristics of included studies

Descriptive information about the studies included within the meta-analyses is detailed in Table 1.

Most studies, 75%, were conducted in the United States of America (k studies = 33), with 25% conducted elsewhere (k = 11). Thirty-eight studies were published between 1970 and 1999, with only six studies published between 2000 and 2017.

There were 152 intervention arms across the 44 studies, of which 91 were active intervention conditions and 61 were control conditions. BTs were most frequently examined (across 59 conditions, k = 33), followed by CBTs (14 conditions, k = 11), SST (10 conditions, k = 10), and Combined interventions (8 conditions, k = 4). Of the control conditions, no treatment control (NTC) was most frequently used (across 29 conditions, k = 29), followed by active control conditions designed to resemble placebo conditions (17 conditions, k = 17), and waitlist control (WLC; 15 conditions, k = 15). Active control conditions included instructional control, non-directive supportive counselling, and interventions with the purportedly 'active' ingredients removed such as implosive treatment based upon neutral cues rather than TA specific cues. No pharmacological interventions were evaluated in any of the included studies (and only one study was identified in our searches but which did not meet our eligibility criteria).

Group format were used most often (76 conditions, k = 34), followed by individual interventions (14 conditions, k = 10), and computer-based delivery (one condition, k = 1). The overall duration of the interventions varied considerably, ranging from one hour to 26 hours (M = 7.24 h, SD = 4.78). Although most studies (k = 30, 68%) did report using cut-off points for TA, this varied greatly from study-to-study, with studies enrolling participants based on: (i) scoring above a cut-off score (k = 19), (ii) scoring above a set percentile for TA severity (ranging from 50th to 85th percentile across studies) (k = 9), (iii) scoring within a range of scores for TA severity (k = 1), or (iv) meeting diagnostic criteria for DSM-IV (American Psychiatric Association, 1994) specific phobia or social phobia (k = 1). Participants self-selected into studies that did not use cut-off criteria (k = 14) (by responding to announcements or advertisements regarding TA trials).

Reporting of the flow of participants within trials was poor; attrition was only reported in 27 of the 44 studies (2, 5–6, 11–15, 17, 19–23, 25, 27, 30–34, 36, 38–41, 44). The mean number of participants enrolled into active treatments was 15.8 (SD = 9.8, range 7 to 48), and for control conditions was 17.1 (SD = 11.0, range 7 to 48). Mean attrition for active conditions was 11.3% (SD = 12, range from 0% to 45%), and for control conditions was 7.8% (SD = 10.3, range from 0% to 35%). Only 15 of the 44 (34%) studies reported follow-up data (1, 4, 6–7, 18–19, 21–22, 24–25, 29, 32, 38–39, 26), with a mean follow-up of 9.4 weeks (SD = 6.4, range from 3 to 26 weeks). Three studies (21, 24–25) had a second follow–up period (M = 56 weeks, SD = 6.9). Only two studies (41–42) included a Consolidated Standards of Reporting Trials (CONSORT; Schulz, Altman, Moher, & Grp, 2010) flow diagram.

Ten different TA measures were used across the studies, with the Achievement Anxiety Test–Debilitating (AAT-D; Alpert & Haber, 1960) administered most frequently (k = 16). A single TA measure was administered in 29 studies, with two or more TA measures administered in 15 studies. There was variation in pre-intervention scores in measures across studies (e.g. between 30.74 and 41.06 on AAT-D). However, given the wide range of outcome measures, the lack of normative data, and no diagnostic criteria for TA, it is difficult to interpret severity of pre-treatment test-anxiety. Four studies administered a state measure of anxiety in an examination context. The effect of interventions on improving academic performance was examined across 17 studies, via students' GPA scores (k = 11, studies 3, 9, 16, 21, 23–26, 32, 38, 41), examination scores shortly after treatment (k = 4, studies 2, 8, 12, 37), or both (k = 2, studies 14, 33).

Only 27 studies reported the number of male and female participants that were enrolled. Gender proportion varied considerably, with

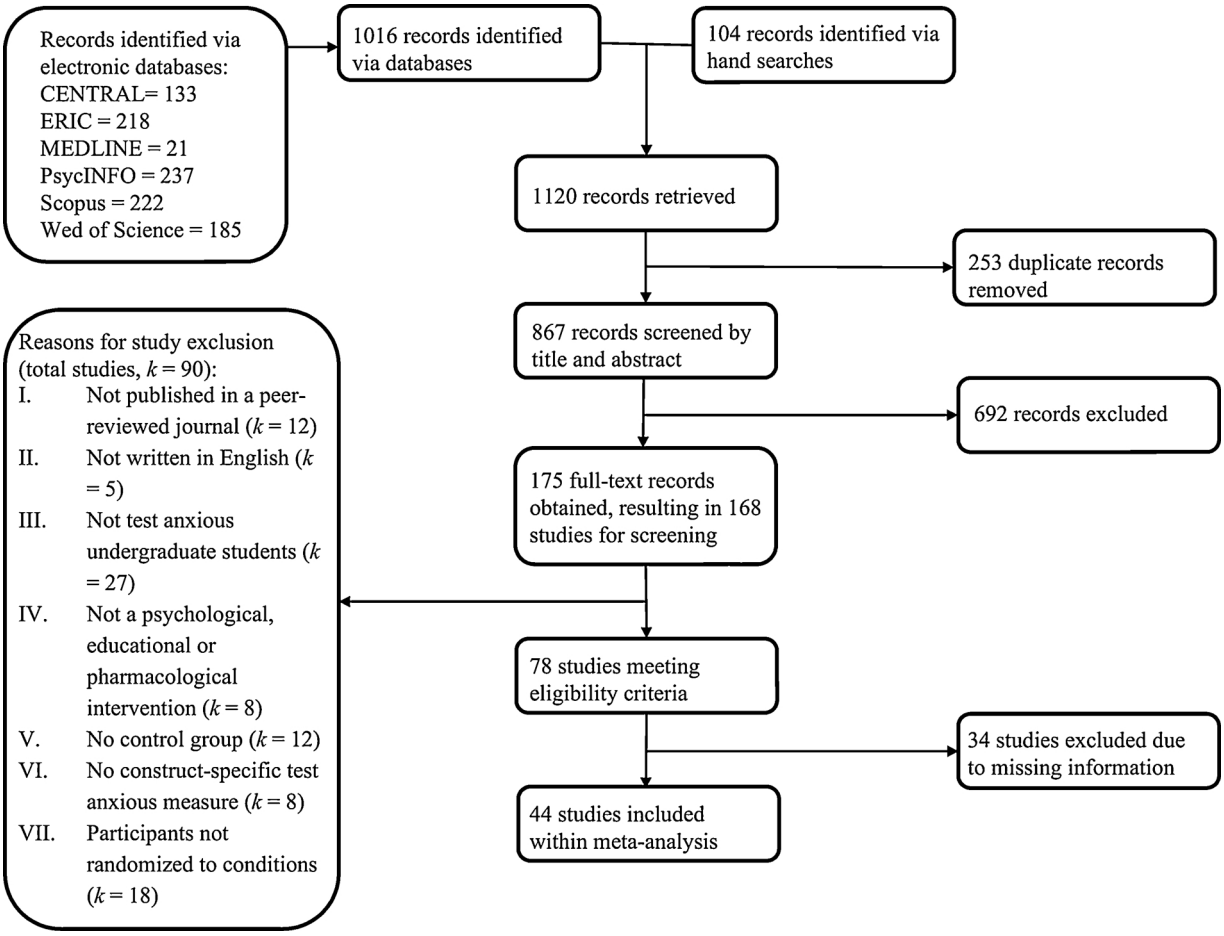

Fig. 1. PRISMA flowchart.

percentages of females ranging from 0% to 97% (M = 64.22, SD = 19.28). Only one study (44) reported the proportion of male and female participants for specific intervention conditions. Just six studies (14%, studies 8, 17, 19, 40, 42, 44) reported the mean age of the sample, and only two (5%, studies 40, 44) reported mean age for specific intervention conditions.

Finally, we also examined the manuscripts to see: (i) if any adverse events had been reported, and (ii) if the clinical significance of the results had been evaluated. No adverse events were reported, nor was the clinical significance of study results evaluated.

3.3. Risk of bias assessment

Fig. 2 summarizes the risk of bias assessments across all included studies. As all included trials used psychological and educational interventions, the blinding of participants items was not included. Overall, 79% of items were scored as 'unclear' indicating poor quality of reporting. Only two studies reported how the random sequence was generated (4.5%), and only one study reported use of allocation concealment (2.3%). Five studies used independent personnel to assess study outcomes (11.4%). Twenty-seven studies (61.4%) were considered to have a low risk of bias with regard to incomplete outcome data. No study made reference to a trial protocol. Inter-rater reliability between the independent raters (JT and ML) was good (Kappa = 0.67). Complete risk of bias assessments for each study can be found in Appendix B (online only).

3.4. Meta-analyses

3.4.1. Categorisation of interventions

Interventions were categorized as follows: (i) BT: anxiety management training, applied relaxation, biofeedback, covert modelling, covert rehearsal, covert reinforcement, cue controlled relaxation, electromyographic-based interventions, eye movement desensitization reprogramming, flooding, imagined modelling, implosive therapy, negative practice, progressive muscle relaxation, relaxation therapy, modelling, self-control desensitization, and systematic desensitization; (ii) CBT: cognitive therapy, cognitive-behavioural therapy (including with imagery rescripting), attention training, and rational emotive therapy; and (iii) Control conditions: attentional control, instructional control, no treatment control, self-help, supportive counselling, and waitlist control. There were no distinct variations of SST. Combined interventions comprised a form of BT or CBT plus SST.

3.4.2. Efficacy of interventions for reducing TA, at post-treatment

From the 44 studies included, 91 active-to-control group comparisons were included within the analyses of post-treatment data. A total of 2209 participants were included in the analyses, with 1362 participants in active treatment conditions and 847 participants in control conditions. Overall, interventions for TA were superior to control conditions in reducing TA severity (see Fig. 3; see also Appendix C, Fig. C.1 for annotated forest plot), with an overall standardized mean effect (Hedges' g) across all interventions of -0.76 (95% CI −0.93 to −0.58).

As expected, significant homogeneity was indicated (I^2 = 66%), and we therefore conducted a subgroups analysis. Combined interventions (g = −1.38, 95% CI −1.96 to −0.81, p < .001, k = 8), BT

Table 1
Selected characteristics of included studies.

ID	First author – year Country	Intervention(s)	Delivery format	Sessions	TA criteria used to select participants	Primary outcome measure	Follow-up period
#1	Mitchell and Ingham (1970) AUS	Systematic Desensitization Waitlist Control No Treatment Control	Group – –	10 × 1 h (twice weekly) – –	≥1 SD above M on AAT-D	AAT-D	14 weeks
#2	Lomont and Sherman (1971) USA	Systematic Desensitization Supportive Counselling No Treatment Control	Group Group –	8 × 1 h (weekly) 8 × 1 h (weekly) –	> 30 on AAT-D & < 23 on AAT-F	AAT-D	–
#3	Prochaska (1971) USA	Implosive Therapy – TA Symptoms Implosive Therapy – TA Dynamic Implosive Therapy – General Anxiety Instructional Control No Treatment Control	Group Group Group Group –	3 × 1 h (fortnightly) 3 × 1 h (fortnightly) 3 × 1 h (fortnightly) 3 × 1 h (fortnightly) –	None	AAT-D	–
#4	Meichenbaum (1972) CAN	Cognitive Therapy Systematic Desensitization Waitlist Control	Group Group –	8 × 1 h (weekly) 8 × 1 h (weekly) –	None	AAT-D	4 weeks
#5	Dawley and Wenrich (1973) USA	Implosive Therapy Instructional Control No Treatment Control	Group Group –	5 × 0.5 h (twice weekly) 5 × 0.5 h (twice weekly) –	≥212 on TAQ (66th percentile)		–
#6	Kostka and Galassi (1974) USA	Systematic Desensitization Covert Modelling No Treatment Control	Group Group. –	10 × 1 h (twice weekly) 8 × 1 h (twice weekly) –	NR	STABS	20 weeks
#7	Guidry and Randolph (1974) USA	Covert Reinforcement Instructional Control No Treatment Control	Group Group –	1 h + 5 × 0.5 h 1 h + 5 × 0.5 h	None	STABS	3 weeks
#8	Mitchell et al. (1975) AUS	Systematic Desensitization + Study Skills Training Relaxation Therapy + Study Skills Training Study Skills Training Instructional Control No Treatment Control	Group Group Group Group –	26 × 1 h (twice weekly) 26 × 1 h (twice weekly) 26 × 1 h (twice weekly) 5 × 1 h (twice weekly) –	None	AAT-D	–
#9	Anton (1976) USA	Systematic Desensitization Supportive Counselling No Treatment Control	Group Group –	8 × 1 h (weekly) 4 × 1 h (weekly) –	None	STAI-T	–
#10	Bedell (1976) USA	Systematic Desensitization – High Systematic Desensitization – Low Relaxation Therapy – High Relaxation Therapy – Low No Treatment Control	Group Group Group Group –	7 × 1 h (weekly) 7 × 1 h (weekly) 7 × 1 h (weekly) 7 × 1 h (weekly) –	NR	TAS	–
#11	Chang-Liang and Denney (1976) USA	Systematic Desensitization Applied Relaxation Relaxation Therapy No Treatment Control	Group Group Group –	3 × 1 h (weekly) 3 × 1 h (weekly) 3 × 1 h (weekly) –	≥66th percentile	STABS	–
#12	Melnick and Russell (1976) USA	Systematic Desensitization Attentional Control No Treatment Control	Group Group –	4 × 1 h (weekly) 4 × 1 h (weekly) –	≥50th percentile on TAQ	TAQ	–
#13	Russell et al. (1976) USA	Systematic Desensitization Cue Controlled Relaxation No Treatment Control	Group Group –	5 × 1 h (weekly) 5 × 0.5 h (weekly) –	None	TAS	–
#14	Denney and Rupert (1977) USA	Self-Control Desensitization–Active Self-Control Desensitization–Passive Systematic Desensitization–Active Systematic Desensitization–Passive Instructional Control No Treatment Control	Ind. Ind. Ind. Ind. Ind. –	8 × 1 h (weekly) 8 × 1 h (weekly) 10 × 1 h (weekly) 10 × 1 h (weekly) 8 × 1 h (weekly) –	≥85th percentile on TAQ	STABS	–
#15	Finger and Galassi (1977)	Attention Training	Group	8 × 1 h (twice weekly)	≥66th	AAT-D	–

(*continued on next page*)

Table 1 (*continued*)

ID	First author – year Country	Intervention(s)	Delivery format	Sessions	TA criteria used to select participants	Primary outcome measure	Follow-up period
	USA	Relaxation Therapy	Group	8 × 1 h (twice weekly)	percentile on		
		Attention Training + Relaxation Therapy	Group	8 × 1 h (twice weekly)	AAT-D-		
		Waitlist Control	–	–	AAT-F scores		
#16	Horne and Matson (1977)	Systematic Desensitization	Group	10 × 1 h (twice weekly)	100 highest	TAS	–
	USA	Flooding	Group	10 × 1 h (twice weekly)	scorers on		
		Modelling	Group	10 × 1 h (twice weekly)	TAS (out of		
		Study Skills Training	Group	10 × 1 h (twice weekly)	175 screened)		
		No Treatment Control	–	–			
#17	Counts, Hollandsworth, and Alcorn (1978)	Electromyographic + Cue Controlled Relaxation	Ind.	6 × 1 h (thrice weekly)	≥21 on TAS	TAS	–
	USA	Cue Controlled Relaxation	Ind.	6 × 1 h (thrice weekly)			
		Attentional Control	Ind.	6 × 1 h (thrice weekly)			
		No Treatment Control	–	–			
#18	Gallagher and Arkowitz (1978)	Covert Modelling	Ind.	4 × 0.5 h (twice weekly)	≥21 on TAS	TAS	3 weeks
	USA	Imagined Modelling	Ind.	4 × 0.5 h (twice weekly)			
		Waitlist Control	–	–			
#19	Lent and Russell (1978)	Study Skills Training + Systematic Desensitization	Group	5 × 1 h (weekly)	Top 30% on	TAS	6 weeks
	USA	Study Skills Training + Cue Controlled Desensitization	Group	5 × 1 h (weekly)	TAS & SSHA		
		Study Skills Training	Group	5 × 1 h (weekly)			
		No Treatment Control	–	–			
#20	Romano and Cabianca (1978)	Electromyographic feedback + Systematic Desensitization	Group	13 × 0.5 h (twice weekly)	≥70th	STABS	–
	USA	Electromyographic feedback	Group	13 × 0.5 h (twice weekly)	percentile		
		Systematic Desensitization	Group	13 × 0.5 h (twice weekly)	on STABS		
		Waitlist Control	–	–			
#21	Deffenbacher, Mathis, and Michaels (1979)	Systematic Desensitization	Group	7 × 1 h (weekly)	> 34 on	AAT-D	6 weeks
	USA	Self-Control Relaxation	Group	7 × 1 h (weekly)	AAT-D		52 weeks
		Waitlist Control	–	–			
		No Treatment Control	–	–			
#22	Deffenbacher and Parks (1979)	Systematic Desensitization	Group	8 × 1 h (twice weekly)	> 25 on	TAS	9 weeks
	USA	Self-Control Desensitization	Group	8 × 1 h (twice weekly)	AAT-D		
		Waitlist Control	–	–			
#23	Holahan, Richardson, Puckett, and Bell (1979)	Cognitive Therapy	Group	4 × 2 h (weekly)	Top 20% on	TAS	–
	USA	Anxiety Management Training	Group	4 × 2 h (weekly)	TAS		
		Waitlist Control	–	–			
#24	Deffenbacher, Michaels, Daley, and Michaels (1980a)	Anxiety Management Training	Group	6 × 1 h (weekly)	Top 10% on	AAT-D	6 weeks
	USA	No Treatment Control	–	–	AAT-D		52 weeks
#25	Deffenbacher, Michaels, Michaels, and Daley (1980b)	Anxiety Management Training	Group	6 × 1 h (weekly)	> 37 on	AAT-D	6 weeks
	USA	Self-Control Desensitization	Group	6 × 1 h (weekly)	AAT-D		64 weeks
		Waitlist Control	–		(Top 5%)		
		No Treatment Control	–	–			
#26	Levine and Obrien (1980)	Negative Practice	Group	6 × 1 h (weekly)	> mean on	TAQ	–
	USA	Negative Practice + Homework	Group	6 × 1 h (weekly)	TAQ		
		Systematic Desensitization	Group	6 × 1 h (weekly)			
		Attentional Control	Group	6 × 1 h (weekly)			
		Waitlist Control	–	–			
#27	Lurie and Steffen (1980)	Covert Reinforcement	Group	6 × 1 h (twice weekly)	≥18 on TAS	TAS	–
	USA	Covert Rehersal	Group	6 × 1 h (twice weekly)			
		Study Skills Training	Group	6 × 1 h (twice weekly)			
		Attentional Control	Group	6 × 1 h (twice weekly)			
		Waitlist Control	–	–			
#28	Reed and Saslow (1980)	Biofeedback	Ind.	8 × 0.5 h (twice weekly)	≥31 on	AAT-D	–
	USA	Instructional Control	Ind.	8 × 0.5 h (twice weekly)	AAT-D		

(*continued on next page*)

Table 1 (*continued*)

ID	First author – year Country	Intervention(s)	Delivery format	Sessions	TA criteria used to select participants	Primary outcome measure	Follow-up period
		No Treatment Control	–	–			
#29	Altmaier and Woodward (1981)	Systematic Desensitization	Group	6 × 1 h (twice weekly)	≥50 on	TAI	12 weeks
	USA	Study Skills Training	Group	6 × 1 h (twice weekly)	TAI		
		Systematic Desensitization + Study Skills Training	Group	12 × 1 h (twice weekly)			
		No Treatment Control	–	–			
#30	Barabasz and Barabasz (1981)	Rational Emotive Therapy	Group	4 × 1 h (twice weekly)	High scorers	TAQ	–
	NZ	Study Skills Training	Group	4 × 1 h (twice weekly)	on skin		
		No Treatment Control	–	–	conductance		
#31	D'Alelio and Murray (1981)	Cognitive Therapy–8	Group	8 × 1.5 h (weekly)	≥55 on TAI	TAI	–
	USA	Cognitive Therapy–4	Group	4 × 1.5 h (weekly)			
		No Treatment Control	–	–			
#32	Decker and Russell (1981)	Cognitive-Behavioural Therapy	Group	4 × 1.5 h (weekly)	≤ mean of	TAS	10 weeks
	USA	Study Skills Training	Group	4 × 1.5 h (weekly)	TAS		
		Waitlist Control	–	–			
#33	McCordick et al. (1981)	Cognitive-Behavioural Therapy + Study Skills Training	Group	15 × 1 h (twice weekly)	None	AAT-D	–
	USA	Cognitive Therapy + Study Skills Training	Group	15 × 1 h (twice weekly)			
		Systematic Desensitization + Study Skills Training	Group	15 × 1 h (twice weekly)			
		Study Skills Training	Group	1 × 1 h (weekly)			
		Waitlist Control	–	–			
#34	Russell and Lent (1982)	Systematic Desensitization	Group	5 × 1 h (weekly)	None	TAS	–
	USA	Cue Controlled Relaxation	Group	5 × 1 h (weekly)			
		Systematic Desensitization + Cue Controlled Relaxation	Group	5 × 1 h (weekly)			
		Attentional Control	Group	5 × 1 h (weekly)			
		No Treatment Control	–	–			
#35	Ricketts and Galloway (1984)	Progressive Muscle Relaxation	Group	1 × 1 h	≥ 147 on	STABS	–
	USA	Study Skills Training	Group	1 × 1 h	STABS		
		Rational Emotive Therapy	Group	1 × 1 h			
		Instructional Control	Group	1 × 1 h			
#36	Crowley, Crowley, and Clodfelter (1986)	Cognitive Therapy	Group	1 × 6 h	≥32 on AAT	TAI	–
	USA	Cognitive Therapy	Ind.	6 × 1 h (twice weekly)			
		No Treatment Control	–	–			
#37	Bauman and Melnyk (1994)	Eye Movement Desensitization Reprogramming	Ind.	1 × 1 h	TAI ≥ 50th	TAI	–
	CAN	No Treatment Control	Ind.	1 × 1 h	percentile		
#38	Sapp (1996)	Relaxation Therapy	Ind.	8 × 1 h (weekly)	None	TAI-W	8 weeks
	USA	Supportive Counselling	Ind.	8 × 1 h (weekly)			
#39	Maxfield and Melnyk (2000)	Eye Movement Desensitization Reprogramming	Ind.	1 × 1.5 h	≥50 on TAI	TAI	8 weeks
	CAN	Waitlist Control	–	–			
#40	Orbach, Lindsay, and Grey (2007)	Cognitive-Behavioural Therapy	Comp.	6 × 0.5 h (weekly)	None	TAI	–
	UK	No Treatment Control	–	–			
#41	Rajiah and Saravanan (2014)	Progressive Muscle Relaxation + Systematic Desensitization	Ind.	6 × 1 h (twice weekly)	None	WTAS	–
	MAL	No Treatment Control	Ind.	–			
#42	Saravanan and Kingston (2014)	Progressive Muscle Relaxation + Systematic Desensitization	Ind.	5 × 1 h (twice weekly)	Score of 30–39	WTAS	–
	MAL	Waitlist Control	–	–	on WTAS		
#43	Hahm, Augustin, Bade, Ammer-Wies, and Bahramsoltani (2016)	Study Skills Training	Group	14 × 1 h (weekly)	None	G-TAI	–
	GER	No Treatment Control	–	–			
#44	Reiss et al. (2017)	Cognitive-Behavioural Therapy	Group	5 × 3 h (weekly)	DSM-IV	G-TAI	26 weeks
	GER	Cognitive-Behavioural Therapy + Imagery Re-scripting	Group	5 × 3 h (weekly)	SCID-I (social		

(*continued on next page*)

Table 1 (*continued*)

ID	First author – year Country	Intervention(s)	Delivery format	Sessions	TA criteria used to select participants	Primary outcome measure	Follow-up period
		Self-help	Group	5 × 3 h (weekly)	/specificphobia)		

Notes. AAT-D/F = Achievement Anxiety Test – Debilitating/Facilitating; Comp = Computer/internet delivered; DSM-IV = Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition; G-TAI = German language version of TAI; h = hour(s); NR = Not Reported; RTAS = Revised Test Anxiety Scale; SCID-I = Structured Clinical Interview DSM-IV Axis I Disorders; SSHA = Survey of Study Habits and Attitudes; STABS = Suinn Test Anxiety Behaviour Scale; TA = Test Anxiety; TAI = Test Anxiety Inventory; TAI-W = TAI–Worry (subscale); TAQ = Test Anxiety Questionnaire; TAS = Test Anxiety Scale; WTAS = Westside Test Anxiety Scale.

($g = -0.83$, 95% CIs -1.06 to -0.60, $k = 59$, $p < .001$), and CBT (g = -0.58, 95% CIs -0.83 to -0.33, $p < .001$, $k = 9$) were significantly superior to control conditions. SST delivered alone ($g = 0.02$, 95% CIs -0.25 to 0.28, $p = .900$, $k = 10$) was not significantly different from control conditions.

Inspection of the forest plot identified three outliers: two in the BT subgroup (studies 41–42) and one in the Combined approach (study 8; SD + SST intervention). Removal of these studies resulted in the overall effect estimate decreasing to $g = -0.64$ (95% CIs -0.77 to -0.50), the BT subgroup effect size changing to $g = -0.70$ (95% CIs -0.88 to -0.52), and the Combined subgroup estimate changing to $g = -1.14$ (95% CIs -1.62 to -0.66).

3.4.3. Efficacy of interventions for reducing TA, at follow-up

Fifteen studies reported follow-up data collection. However, data were only available for 12 studies (studies 1, 4, 6–7, 21–22, 24–25, 32, 38–39, 44), with three studies not reporting sufficient data to be included in the analyses (studies 18–19, 29). We included only data from the first follow-up period (studies 21, 24–25 had a second follow-up period). Data from 294 participants across 19 treatment arms/conditions and 239 participants in control conditions were included in our analyses. Of the active intervention conditions, data were available for BT (14 conditions, $k = 9$), CBT (4 conditions, $k = 3$), and SST (1 condition, $k = 1$). Overall, interventions for TA were superior to control conditions at follow-up ($g = -0.97$, 95% CI -1.31 to -0.64, $p < .001$) (see Appendix C for forest and funnel plots). Significant heterogeneity was indicated ($I^2 = 64\%$). Subgroups analysis showed that BT ($g = -1.12$, 95% CI -1.47 to -0.77, $p < .001$, $k = 9$) and CBT ($g = -0.40$, 95% CI -0.92 to 0.12, $p = .007$, $k = 3$) were superior to control conditions. We were unable to compute estimates for the other subgroups (SST, Combined) due to insufficient data. Inspection of the forest plot revealed no outliers.

3.4.4. Efficacy of interventions for reducing in situ (state) TA

Data from 10 conditions ($k = 3$) were analyzed (from studies 13, 16, 32). In total, data from 128 participants across 10 treatment arms/conditions and 36 participants in control conditions were included in our analyses. Overall, interventions were superior to control conditions in reducing in situ (state) TA experienced immediately prior to 'real world' examinations ($g = -1.20$, 95% CI -1.66 to -0.75, $p < .001$, $k = 3$). Significant heterogeneity was not indicated ($I^2 = 18\%$). No outliers were identified from the forest plot (see Appendix C for forest and funnel plots).

3.4.5. Efficacy of interventions for improving academic performance, at post-treatment

Data from 36 treatment arms ($k = 17$) were analyzed (from studies 2–3, 8–9, 12, 14, 16, 21, 23–26, 32–33, 37–38, 41), which comprised 25 treatment arms/conditions within the BT category ($k = 14$), five within Combined ($k = 2$), four within SST ($k = 4$), and two conditions for CBT ($k = 2$). In total, data from 548 participants across active treatment arms/conditions and 363 participants in control conditions were included in our analyses. Within our analyses, results from single examinations came from 10 treatment arms/conditions ($k = 4$), while GPA results came from 26 treatment arms/conditions ($k = 13$). Overall, treatments were superior to control conditions in improving academic performance producing a small-to-moderate effect ($g = 0.37$, 95% CI 0.14 to 0.61, $p = .002$, $k = 36$) (see Appendix C for forest and funnel plots). Significant heterogeneity was indicated ($I^2 = 61\%$). Subgroups analysis showed that BT ($g = 0.22$, 95% CI 0.05 to 0.40, $p = .021$, $k = 13$) and Combined interventions ($g = 1.58$, 95% CI 0.41 to 2.76, $p < 001$, $k = 2$) were superior to control conditions, but SST ($g = 0.34$, 95% CI -0.16 to 0.84, $p = .790$, $k = 4$) and CBT ($g = -0.24$, 95% CI -0.98 to 0.49, $p = .520$, $k = 4$) were not significantly different from control conditions. Inspection of the forest plot revealed one outlier in the Combined subgroup (systematic desensitization and SST

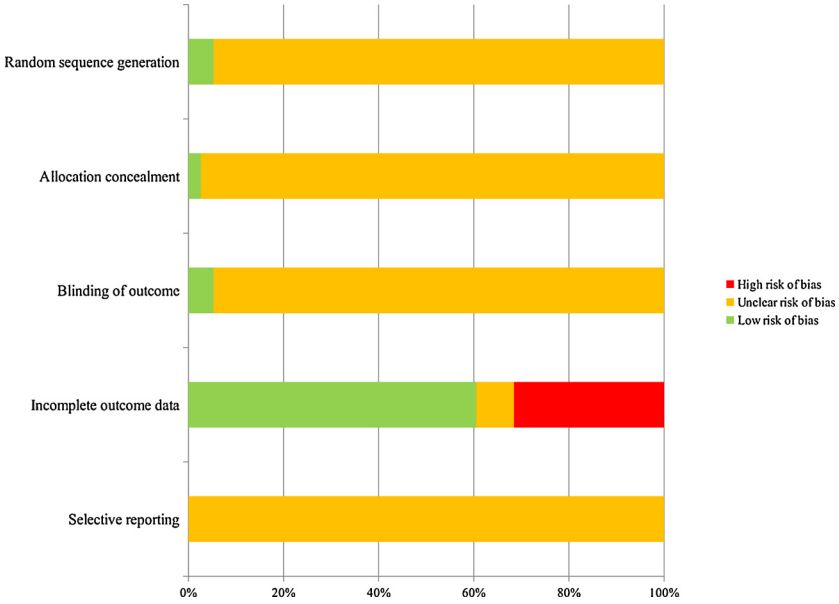

Fig. 2. Summary of risk of bias assessments.

intervention of Mitchell, Hall, & Piatkowska, 1975). Removal of this outlier resulted the overall effect estimate changing to $g = 0.28$ (95% CIs 0.11 to 0.46) and the Combined interventions subgroup effect size changing to $g = 1.15$ (95% CIs 0.33 to 1.96).

3.4.6. Efficacy of interventions for improving academic performance, at follow-up

Four studies reported follow-up data collection (studies 14, 21, 23, 38). We included only data from the first follow-up period (one study had a second follow-up period). It was not possible to calculate the mean follow-up period as insufficient data was provided. In total data from 121 participants across nine treatment arms/conditions and 84 participants in control conditions were included in our analyses. Of the active intervention conditions, data were available for BT (8 conditions, $k = 4$) and CBT (1 condition, $k = 1$). Overall, interventions for TA were not significantly different to control conditions at follow-up ($g = 0.23$, 95% CI −0.11 to 0.56, $p = .180$) (see Appendix C for forest and funnel plots, online only). Significant heterogeneity was not indicated ($I^2 = 25\%$). No outliers were identified.

A summary of effect sizes across intervention approaches is in Table 2.

3.4.7. Moderators of intervention effect

We examined moderators of treatment outcome for BT interventions at post-treatment for TA reduction only, as there was insufficient data to examine moderators for the other treatment approaches (and insufficient data for all approaches at follow-up). Significant heterogeneity was indicated within outcomes for specific BT interventions ($I^2 = 70\%$). The two sample-dependent moderators – gender, pre-treatment TA severity – had to be dropped from our analyses due to insufficient data. Thus, only format of the intervention (group vs. individual), dosage (number of treatment hours), and manualization (poor vs. adequate) were entered into the meta-regression; none were significant moderators of treatment outcome. Results from the multiple meta-regression analyses are presented in Appendix D (online only).

3.4.8. Sensitivity analyses

Sensitivity analyses were planned to evaluate the robustness of the results by removing studies with a high risk of bias and comparing the results when all studies are included. However, these analyses could not be conducted as the poor quality of reporting was such that the overwhelming majority of risk of bias domains were classified as unclear (see Fig. 2) and so no RCT could be considered of high (or low) risk of bias.

3.4.9. Assessment of publication bias

Publication bias was explored using data derived from TA interventions at post-treatment. Inspection of the funnel plot (see Appendix C, Fig. C.2) revealed minor asymmetry and this was confirmed by Egger's regression test ($z = -2.11$, $p = .003$). More noticeable from the funnel plot was the absence of studies with very precise estimates (i.e., that had sufficiently large samples) included within the analysis.

3.5. Quality of evidence

GRADE assessments were made for each intervention subgroup across five outcomes: reducing TA at (i) post-treatment, (ii) follow-up; (iii) reducing in situ (state) TA experienced immediately prior to sitting a real examination; and improving academic performance at (iv) post-treatment, and (v) follow-up (see Appendix E, online only). For a GRADE assessment to be made for an intervention approach on any of the above outcomes there must be evidence from at least two independent studies. The overall quality of evidence was poor, with 'low quality' being the highest grading awarded for any intervention on any outcome. The main reasons for the low quality ratings were: (i) potential for risk of bias, indicated by nearly 80% of ratings being scored 'unclear', (ii) inconsistency as indicated by I^2 scores, and (iii) the relative imprecision of the effect size estimates (i.e., relatively large confidence intervals around the effect size estimate). There were also large gaps in the evidence base for all intervention approaches except for BTs. For reducing TA at post-treatment, the evidence was rated as 'low quality' for all intervention approaches. At post-treatment BT and CBT were both rated as 'low quality', while lack of evidence meant no ratings could be given to SST and Combined intervention approaches. Evidence for reducing in situ (state) TA was rated as 'very low quality' for BT and SST, but no ratings could be made for CBT and Combined approaches. With regard to improving academic performance at post-treatment, evidence for BT, CBT, SST was rated as 'low quality', while evidence for Combined interventions approaches were rated as 'very low quality'. For improving academic performance at follow-up, the evidence for BT was rated as 'very low quality'; there was insufficient evidence for the other approaches to award a rating. Overall, these

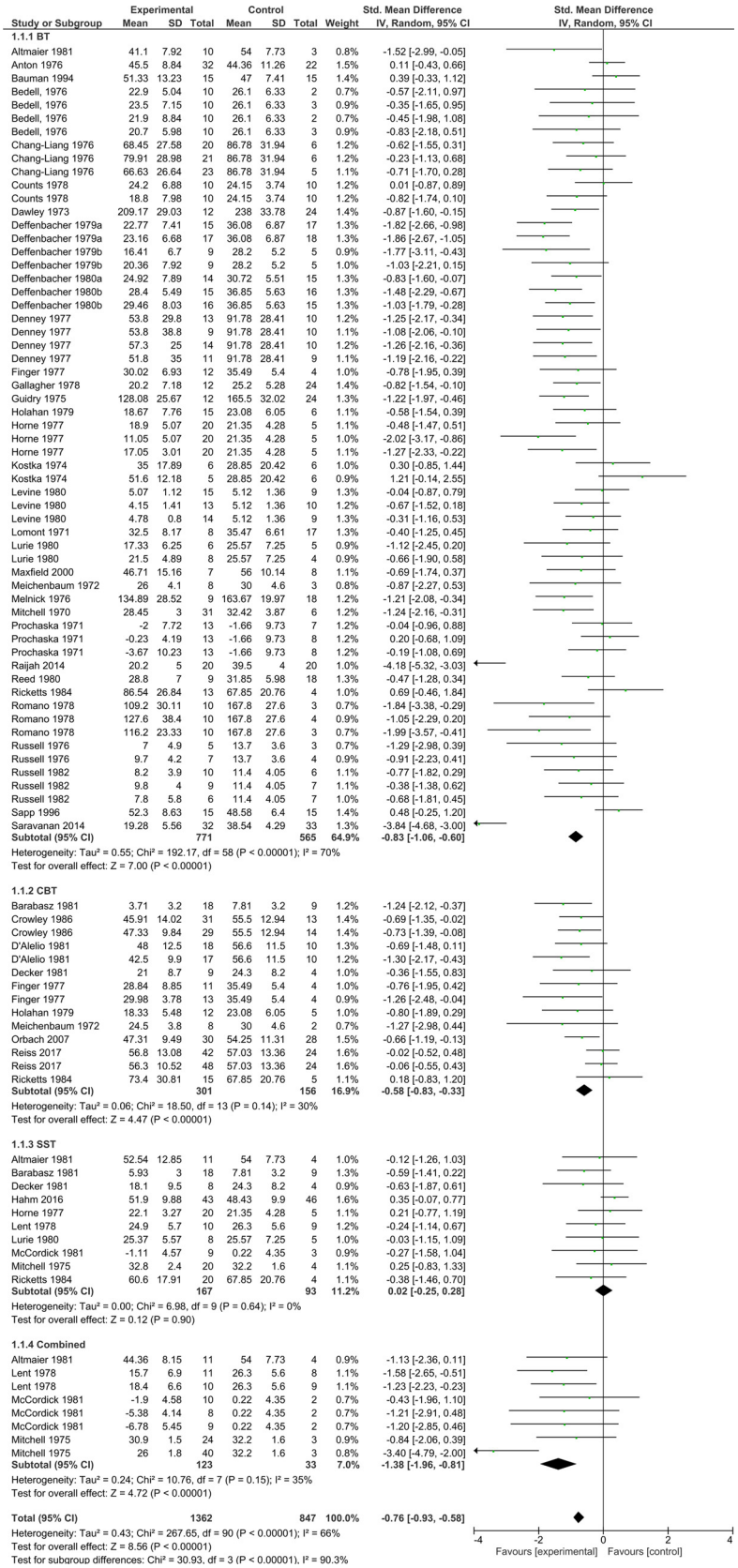

Fig. 3. Forest plot of efficacy of interventions for reducing test anxiety at post-treatment (compared to control conditions).

Table 2
Summary of meta-analytic findings (active interventions vs. control conditions).

	Test anxiety			Academic performance	
Intervention	Post	Follow-up	State	Post	Follow-up
BT	−0.83 (−1.06, −0.60)	−1.15 (−1.38, −0.91)	−1.52 (−2.06, −0.97)	0.22 (0.05, 0.40)	0.20 (−0.17, −0.56)
CBT	−0.58 (−0.83, −0.33)	−0.31 (−0.64, 0.02)	–	−0.24 (−0.98, 0.49)	0.54 (−0.41, 1.50)
SST	0.02 (−0.25, 0.28)	−0.95 (−2.24, 0.34)	−0.73 (−2.57, 1.12)	0.34 (−0.16, 0.84)	–
Combined	−1.38 (−1.96, −0.81)	–	−0.85 (−1.77, 0.08)	1.58 (0.41, 2.76)	–
Overall	*−0.76 (−0.93, −0.58)*	*−0.87 (−1.06, −0.68)*	*−1.20 (−1.66, −0.75)*	*0.37 (0.14, 0.61)*	*0.23 (−0.11, 0.56)*

Notes. Test anxiety: negative effect sizes indicate reduction in test anxiety; positive effect sizes indicate increase in test anxiety. Academic performance: negative effect sizes indicate decrease in academic performance; positive effect sizes indicate increase in academic performance. BT = Behavioural therapy, CBT = Cognitive-behavioural therapy, Combined = Combined psychotherapy (BT, CBT) and SST; SST = Study Skills Training.

findings indicate that future studies will very likely have significant impact on the effect size estimates and their confidence intervals.

4. Discussion

4.1. Summary of main results

This meta-analysis is the first to systematically examine evidence regarding the efficacy of interventions for test-anxious university undergraduate students. Our aims were to examine the relative efficacy of interventions for reducing TA and improving academic performance. We also aimed to examine possible moderators of intervention effect. Overall, interventions were superior to control conditions at post-treatment for reducing TA, with the summary effect size across all interventions ($g = -0.76$) indicating a moderate-to-large effect (based upon Cohen's 1988 guidelines). However, as expected, there was significant heterogeneity between study outcomes. In a subgroup analysis, combined psychological and SST interventions, BT, and CBT were superior to control conditions, but SST delivered alone was not superior to control conditions. An important question is whether intervention effects were sustained over follow-up (compared with control conditions)? Overall, interventions were superior to control conditions at follow-up, with a large effect size found ($g = -0.97$). However, the length of the follow-up period across the 12 studies that reported data was relatively brief (at a mean of 9.4 weeks) and did not specify when follow-up assessment took place in relation to the timing of examinations. Only BT was superior to control conditions at follow-up.

Given TA is a situation-specific form of anxiety, we explored whether interventions had a beneficial impact on anxiety experienced immediately before an examination (in situ state anxiety). A large overall effect size was found ($g = -1.20$) indicating interventions may help alleviate TA immediately prior to an examination. However, only the effect size estimate was based on data from just three studies.

Interventions also improve academic performance. Interventions were significantly superior control conditions ($g = 0.37$) at post-treatment. Subgroups analysis found BT and Combined psychological and educational intervention approaches superior to control conditions but SST delivered alone was not superior to control conditions. There was insufficient data available to calculate an effect for CBT. At follow-up, no significant differences between active interventions and control conditions were found, though data were derived from just four studies.

The moderation analysis of BT effect sizes at post-treatment, which examined delivery format (group vs. individual), dosage (number of contact hours), and whether the intervention was manualized, found all were non-significant predictors. There was insufficient data for moderator analyses to be conducted for the other interventions and follow-up periods.

Although there were generally large effect sizes for treatments compared to control conditions, the overall quality of evidence was poor – or insufficient details were provided by the individual RCTs meaning GRADE assessments could not be made. Across the intervention approaches, BT had the most substantive evidential support. However, even within the BT approach there was high heterogeneity that suggested different levels of efficacy between specific behavioural treatments. There is less evidence for CBT with moderate post-treatment effects but limited follow-up data negates assessment of longer-term effects. SST interventions do not appear effective for reducing TA nor are particularly effective at improving academic performance. However, SSTs interventions included were generally poorly described, and it might be that manualized educational interventions, informed by recent literature in this field (e.g., retrieval-practice effect; Karpicke & Roediger, 2008) would deliver superior outcomes. It should also be noted that the primary aim of the SST interventions was to decrease TA with improvement in academic performance a secondary aim, whereas most educational interventions aim to improve academic performance first with the impact on TA rarely considered. Combined psychological and SST interventions seem promising. However, a lack of data meant an effect size could only be calculated at post-treatment, so the longer-term efficacy of these approaches is unknown.

4.2. Overall completeness and applicability of evidence

Studies included in our review are a partial representation of the studies conducted and published. We identified 78 eligible studies. Thirty four of these studies could not be included within our meta-analyses as insufficient data were reported in the study manuscripts and, when contacted, authors no longer had access to the data. Additionally, publication bias was indicated by marginal funnel plot asymmetry, suggesting some trials with poorer outcomes have not been published in peer-reviewed journals. Science is a collaborative and cumulative endeavour and efforts should be made in the future to catalog and preserve such trial data in the future (Goldacre, 2015).

All reviewed studies included university undergraduates only but important demographic and clinical information was frequently not reported. Only 27 of the 44 studies reported the percentage of male and female participants in their samples, with the ratio of male-to-female participants varying widely, from 0% (Prochaska, 1971) to 97% (Bauman & Melnyk, 1994). Just one study reported the proportion of male and female participants for intervention conditions (Reiss et al., 2017). These figures cast doubt on the representativeness of the samples used in some studies.

Reporting of the clinical characteristics of study samples was similarly incomplete. Only 30 of the studies used cut-off criteria, with the other 14 studies just enrolling self-selected volunteers who responded to advertisements. The most frequent method was to screen an entire year's cohort and offer participation in the trial to those scoring above a cut-score or set percentile. However, there were large discrepancies in how the cut-offs were applied. For example, eligible participants could come from the top 50% of scores (Melnick & Russell, 1976) to the top 15% (Denney & Rupert, 1977). Only one study (Reiss et al., 2017) used a structured clinical interview to screen for TA. Thus, the severity of TA varied widely from study-to-study.

4.3. Quality of evidence

Overall, results should be interpreted with caution due the issues we

identified. The principal problem was the poor quality of reporting, with many details either partially reported or not reported at all. For example, only two studies reported the flow of participants from recruitment to follow-up periods (Rajiah & Saravanan, 2014; Reiss et al., 2017). Nearly 80% of risk of bias items were scored as unclear due to insufficient information reported in study manuscripts and we cannot be confident those trials were conducted rigorously. This meant it was not possible to identify any high (or low) quality trials. Poor quality reporting does not necessarily mean methodologically poor trials. Rather, it means one's confidence in the estimates obtained is undermined. Most interventions examined in studies were insufficiently described, with nearly half of studies (42%) not referencing a manual (or how different components were combined). Manualization is critical as it enhances the internal validity of the trial and allows replicability (Temple, Salmon, Tudur-Smith, Huntley, & Fisher, 2018).

The review identified 44 studies. No effect size was estimated with precision for any of the subgroups, with the smallest interval between upper and lower 95% bounds of the effect size estimate being 0.46 (for BT subgroup for reducing TA severity at post-treatment). This is a result of the low statistical power for individual studies, confirmed by visual inspection of the funnel plot, where no studies are located at the apex of the funnel (i.e., this is where studies with small standard errors – indicative of large samples – should be). To detect an effect size of 0.70, with 80% power, and α of .05, a minimum of 34 participants are required for each treatment arm (two-side significance). However, only 5% (8 of 152) treatment arms had 34 or more participants, with a mean of just 15.8 participants allocated to active conditions.

Another limitation of the studies included was the large number of different self-rating scales that were used to assess TA severity. This makes it difficult to compare TA severity. A 'gold standard' outcome measure is needed.

4.4. Limitations of this review

There are several limitations to this review. Firstly, the search protocol was limited to English-language peer-reviewed published studies. Cochrane review protocols suggest attempting to find all non-published and non-English documents (Higgins & Green, 2011), and the inclusion of such studies would have been preferable.

Secondly, in our statistical analyses we dealt with studies with multiple treatment arms by dividing the sample size of the control group by the number of treatment arms. This meant each participant was only included once in each analysis. Other methods for dealing with this include combining treatment arms that are conceptually similar. However, as this was the first review of interventions for TA in undergraduate students, we believed it was better to show the results from each treatment arm, rather than obfuscating what was originally examined by combining interventions. We also decided (a priori) to keep the subgroups relatively broad (e.g., BT) as we anticipated that evidence available for interventions based upon specific models (e.g., Ellis' Rational Emotive Behaviour Therapy; Ellis, 1962) would be limited.

5. Conclusion

BT is the current first choice intervention for reducing TA and, in turn, improving academic performance. There is more substantial evidence for the efficacy of BT immediately following treatment and over the follow-up period compared to other intervention approaches. However, there was inconsistency in effects across the various forms of BT interventions making it difficult to recommend a specific behavioural intervention. Combined interventions (BT or CBT plus SST) are promising but there is lack of evidence for longer-term efficacy. CBT appears to be efficacious for reducing TA severity, but the magnitude of its effects is less than either BT or the Combined approaches. SST is ineffective in alleviating TA in the short-term but has more promising longer-term effects, while it also appears to improve short-term academic performance. In general, the interventions evaluated have a greater impact on TA than academic performance. Moreover, overall confidence in the results should be tempered, given the relatively small number of included studies, evidence of publication bias, relatively wide confidence intervals around effect size estimates, and the poor quality of reporting of individual RCTs.

In summary, higher quality, larger trials are required. These studies should be adequately powered, include longer follow-ups, and adhere to modern reporting standards. Additionally, given that TA is a situation-specific disorder, future trials should also consider when post- and follow-up assessments are made (i.e., ideally soon after examinations) so that the effects of interventions are clearer. Finally, our review highlights a marked drop in the number of trials evaluating interventions designed to alleviate TA. High quality trials of other psychological interventions for TA are needed. For example, Intolerance of Uncertainty therapy (Dugas, Gagnon, Ladouceur, & Freeston, 1998) and metacognitive therapy (Wells, 2000, 2009) aim to modify the psychological mechanisms underpinning worry, the defining feature of TA. These interventions may prove to be more effective in reducing TA and improving academic performance than prior approaches.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.janxdis.2019.01.007.

References[1]

Addis, M. E., Cardemil, E. V., Duncan, B. L., & Miller, S. D. (2006). *Does Manualization Improve Therapy Outcomes? Evidence-based practices in mental health: Debate and dialogue on the fundamental questions.* Washington, DC: American Psychological Association131–160.

Alpert, R., & Haber, R. N. (1960). Anxiety in academic achievement situations. *The Journal of Abnormal and Social Psychology, 61*(2), 207–215. https://doi.org/10.1037/h0045464.

*Altmaier, E. M., & Woodward, M. (1981). Group vicarious desensitization of test anxiety. *Journal of Counseling Psychology, 28*(5), 467–469. https://doi.org/10.1037/0022-0167.28.5.467.

American Psychiatric Association (1994). *Diagnostic and statistical manual for mental disorders.* Washington, DC: Author.

*Anton, W. D. (1976). Evaluation of outcome variables in systematic desensitization of test anxiety. *Behaviour Research and Therapy, 14*(3), 217–224. https://doi.org/10.1016/0005-7967(76)90014-0.

*Barabasz, A. F., & Barabasz, M. (1981). Effects of rational-emotive therapy on psychophysiological and reported measures of test anxiety arousal. *Journal of Clinical Psychology, 37*(3), 511–514. https://doi.org/10.1002/1097-4679(198107)37:3<511::aid-jclp2270370311>3.0.co;2-3.

*Bauman, W., & Melnyk, W. T. (1994). A controlled comparison of eye-movements and finger tapping in the treatment of test anxiety. *Journal of Behavior Therapy and Experimental Psychiatry, 25*(1), 29–33. https://doi.org/10.1016/0005-7916(94)90060-4.

Beck, A. T., Emery, G., & Greenberg, R. L. (1986). *Anxiety disorders and phobias: A cognitive perspective.* New York, NY: Basic Books.

*Bedell, J. R. (1976). Systematic desensitization, relaxation training and suggestion in the treatment anxiety. *Behaviour Research and Therapy, 14*(4), 309–311. https://doi.org/10.1016/0005-7967(76)90008-5.

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

Brewer, C. (1971). Beneficial effect of beta-adrenergic blockade on "exam nerves". *The Lancet, 300*(7774), 435. https://doi.org/10.1016/S0140-6736(72)91840-5.

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology, 27*(2), 270–295. https://doi.org/10.1006/ceps.2001.1094.

*Chang-Liang, R., & Denney, D. R. (1976). Applied relaxation as training in self-control. *Journal of Counseling Psychology, 23*(3), 183–189. https://doi.org/10.1037//0022-0167.23.3.183.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences.* Hillsdale, NJ: Erlbaum.

*Counts, D. K., Hollandsworth, J. G., & Alcorn, J. D. (1978). Use of electromyographic biofeedback and cue-controlled relaxation in treatment of test anxiety. *Journal of Consulting and Clinical Psychology, 46*(5), 990–996. https://doi.org/10.1037//0022-

[1] References marked with an asterisk (*) indicate studies included within the meta-analysis.

006x.46.5.990.
*Crowley, C., Crowley, D., & Clodfelter, C. (1986). Effects of a self-coping cognitive treatment for test anxiety. *Journal of Counseling Psychology, 33*(1), 84–86. https://doi.org/10.1037//0022-0167.33.1.84.
*D'Alelio, W. A., & Murray, E. J. (1981). Cognitive therapy for test anxiety. *Cognitive Therapy and Research, 5*(3), 299–307. https://doi.org/10.1007/bf01193413.
*Dawley, H. H., & Wenrich, W. W. (1973). Treatment of test anxiety by group implosive therapy. *Psychological Reports, 33*(2), 383–388.
*Decker, T. W., & Russell, R. K. (1981). Comparison of cue-controlled relaxation and cognitive restructuring versus study skills counseling in treatment of test-anxious college underachievers. *Psychological Reports, 49*(2), 459–469.
*Deffenbacher, J. L., Mathis, H., & Michaels, A. C. (1979). Two self-control procedures in the reduction of targeted and nontargeted anxieties. *Journal of Counseling Psychology, 26*(2), 120–127. https://doi.org/10.1037//0022-0167.26.2.120.
*Deffenbacher, J. L., Michaels, A. C., Daley, P. C., & Michaels, T. (1980a). A comparison of homogeneous and heterogeneous anxiety management training. *Journal of Counseling Psychology, 27*(6), 630–634.
*Deffenbacher, J. L., Michaels, A. C., Michaels, T., & Daley, P. C. (1980b). Comparison of anxiety management training and self-control desensitization. *Journal of Counseling Psychology, 27*(3), 232–239. https://doi.org/10.1037/0022-0167.27.3.232.
*Deffenbacher, J. L., & Parks, D. H. (1979). A comparison of traditional and self-control desensitization. *Journal of Counseling Psychology, 26*(2), 93–97. https://doi.org/10.1037/0022-0167.26.2.93.
Dendato, K. M., & Diener, D. (1986). Effectiveness of cognitive/relaxation therapy and study-skills training in reducing self-reported anxiety and improving the academic performance of test-anxious students. *Journal of Counseling Psychology, 33*(2), 131–135. https://doi.org/10.1037/0022-0167.33.2.131.
*Denney, D. R., & Rupert, P. A. (1977). Desensitization and self-control in the treatment of test anxiety. *Journal of Counseling Psychology, 24*(4), 272–280. https://doi.org/10.1037//0022-0167.24.4.272.
Depreeuw, E., & De-Neve, H. (1992). Test anxiety can harm your health: Some conclusions based on a student typology. In D. G. Forgays, T. Sosnowski, & K. Wrzesniewski (Eds.). *Anxiety: Recent developments in cognitive, psychophysiological, and health research* (pp. 211–228). Washington: Hemisphere.
Dugas, M. J., Gagnon, F., Ladouceur, R., & Freeston, M. H. (1998). Generalized anxiety disorder: A preliminary test of a conceptual model. *Behaviour Research and Therapy, 36*(2), 215–226. https://doi.org/10.1016/s0005-7967(97)00070-3.
Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*(7109), 629–634.
Ellis, A. (1962). *Reason and emotion in psychotherapy.* New York: Stuart.
Ergene, T. (2003). Effective interventions on test anxiety reduction – A meta-analysis. *School Psychology International, 24*(3), 313–328. https://doi.org/10.1177/01430343030243004.
*Finger, R., & Galassi, J. P. (1977). Effects of modifying cognitive versus emotionality responses in the treatment of test anxiety. *Journal of Consulting and Clinical Psychology, 45*(2), 280–287. https://doi.org/10.1037//0022-006x.45.2.280.
*Gallagher, J. W., & Arkowitz, H. (1978). Weak effects of covert modeling treatment of test anxiety. *Journal of Behavior Therapy and Experimental Psychiatry, 9*(1), 23–26. https://doi.org/10.1016/0005-7916(78)90083-6.
Goldacre, B. (2015). How to get all trials reported: Audit, better data, and individual accountability. *PLoS Medicine, 12*(4), e1001821. https://doi.org/10.1371/journal.pmed.1001821.
*Guidry, L. S., & Randolph, D. L. (1974). Covert reinforcement in the treatment of test anxiety. *Journal of Counseling Psychology, 21*(4), 260–264. https://doi.org/10.1037/h0036724.
Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., ... Grp, G. W. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *British Medical Journal, 336*(7650), 924–926. https://doi.org/10.1136/bmj.39489.470347.AD.
*Hahm, N., Augustin, S., Bade, C., Ammer-Wies, A., & Bahramsoltani, M. (2016). Test anxiety: Evaluation of a low-threshold seminar-based intervention for veterinary students. *Journal of Veterinary Medical Education, 43*(1), 47–57. https://doi.org/10.3138/jvme.0215-029R1.
Hahne, R., Lohmann, R., & Krzyszycha, K. (1999). *Studium und psychische Probleme: Sonderauswertung zur 15 Sozialerhebung des Deutschen Studentenwerks.* Bonn: Deutsches Studentenwerk.
Hedges, L. V. (1981). Distribution Theory for Glass's Estimator of Effect Size and Related Estimators. *Journal of Educational Statistics, 6*(2), 107–128. https://doi.org/10.2307/1164588.
Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58*(1), 47–77. https://doi.org/10.3102/00346543058001047.
Herzer, F., Wendt, J., & Hamm, A. O. (2014). Discriminating clinical from nonclinical manifestations of test anxiety: A validation study. *Behavior Therapy, 45*(2), 222–231. https://doi.org/10.1016/j.beth.2013.11.001.
Higgins, J. P. T., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., ... Sterne, J. A. C. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *British Medical Journal (Online), 343*(7829), https://doi.org/10.1136/bmj.d5928.
Higgins, J. P. T., & Green, S. (Eds.). (2011). *Cochrane handbook for systematic reviews of interventions version 5.1.0*The Cochrane Collaboration.
Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557.
Hill, K. T., & Wigfield, A. (1984). Test anxiety: A major educational problem and what can be done about it. *Elementary School Journal, 85*(1), 105–126. https://doi.org/10.1086/461395.
*Holahan, C. J., Richardson, F. C., Puckett, S. P., & Bell, K. F. (1979). Evaluation of two test-anxiety reduction treatments in a secondary prevention program. *American Journal of Community Psychology, 7*(6), 679–687. https://doi.org/10.1007/bf00891970.
*Horne, A. M., & Matson, J. L. (1977). Comparison of modeling, desensitization, flooding, study skills, and control groups for reducing test anxiety. *Behavior Therapy, 8*(1), 1–8. https://doi.org/10.1016/s0005-7894(77)80114-7.
Huntley, C. D., Young, B., Jha, V., & Fisher, P. L. (2016). The efficacy of interventions for test anxiety in university students: A protocol for a systematic review and meta-analysis. *International Journal of Educational Research, 77*, 92–98. https://doi.org/10.1016/j.ijer.2016.03.001.
Jacobson, E. (1938). *Progressive relaxation* (2nd ed.). Oxford, UK: University of Chicago Press.
Kavakci, O., Semiz, M., Kartal, A., Dikici, A., & Kugu, N. (2014). Test anxiety prevalence and related variables in the students who are going to take the university entrance examination. *Düşünen Adam: The Journal of Psychiatry and Neurological Sciences, 27*, 301–307.
Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319*(5865), 966–968. https://doi.org/10.1126/science.1152408.
Knappe, S., Beesdo-Baum, K., Fehm, L., Stein, M. B., Lieb, R., & Wittchen, H. U. (2011). Social fear and social phobia types among community youth: Differential clinical features and vulnerability factors. *Journal of Psychiatric Research, 45*(1), 111–120. https://doi.org/10.1016/j.jpsychires.2010.05.002.
*Kostka, M. P., & Galassi, J. P. (1974). Group systematic desensitization versus covert positive reinforcement in the reduction of test anxiety. *Journal of Counseling Psychology, 21*(6), 464–468. https://doi.org/10.1037/h0037290.
*Lent, R. W., & Russell, R. K. (1978). Treatment of test anxiety by cue-controlled desensitization and study-skills training. *Journal of Counseling Psychology, 25*(3), 217–224. https://doi.org/10.1037//0022-0167.25.3.217.
*Levine, R. A., & Obrien, R. M. (1980). Treatment of anxiety about college tests with negative practice and systematic desensitization: Some negative findings. *Psychological Reports, 46*(3), 823–829.
*Lomont, J. F., & Sherman, L. J. (1971). Group systematic desensitization and group insight therapies for test anxiety. *Behavior Therapy, 2*(4), 511–518. https://doi.org/10.1016/s0005-7894(71)80097-7.
*Lurie, E. S., & Steffen, J. J. (1980). Effective components of covert reinforcement. *Journal of Psychology, 106*(2), 241–248.
*Maxfield, L., & Melnyk, W. T. (2000). Single session treatment of test anxiety with eye movement desensitization and reprocessing (EMDR). *International Journal of Stress Management, 7*(2), 87–101. https://doi.org/10.1023/a:1009580101287.
*McCordick, S. M., Kaplan, R. M., Smith, S., & Finn, M. E. (1981). Variations in cognitive behavior modification for test anxiety. *Psychotherapy—Theory Research and Practice, 18*(2), 170–178. https://doi.org/10.1037/h0086077.
*Meichenbaum, D. H. (1972). Cognitive modification of test-anxious college students. *Journal of Consulting and Clinical Psychology, 39*(3), 370–380.
*Melnick, J., & Russell, R. W. (1976). Hypnosis versus systematic desensitization in the treatment of test anxiety. *Journal of Counseling Psychology, 23*(4), 291–295. https://doi.org/10.1037//0022-0167.23.4.291.
*Mitchell, K. R., Hall, R. F., & Piatkowska, O. E. (1975). A group program for treatment of failing college students. *Behavior Therapy, 6*(3), 324–336. https://doi.org/10.1016/s0005-7894(75)80107-9.
*Mitchell, K. R., & Ingham, R. J. (1970). The effects of general anxiety on group desensitization of test anxiety. *Behaviour Research and Therapy, 8*(1), 69. https://doi.org/10.1016/0005-7967(70)90037-9.
PRISMA GroupMoher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. . (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*(7), https://doi.org/10.1371/journal.pmed.1000097.
Naveh-Benjamin, M., Lavi, H., McKeachie, W. J., & Lin, Y. G. (1997). Individual differences in students' retention of knowledge and conceptual structures learned in University and High school courses: The case of test anxiety. *Applied Cognitive Psychology, 11*(6), 507–526. https://doi.org/10.1002/(sici)1099-0720(199712)11:6<507::aid-acp482>3.0.co;2-g.
Neuderth, S., Jabs, B., & Schmidtke, A. (2009). Strategies for reducing test anxiety and optimizing exam preparation in German university students: A prevention-oriented pilot project of the University of Wurzburg. *Journal of Neural Transmission, 116*(6), 785–790. https://doi.org/10.1007/s00702-008-0123-7.
*Orbach, G., Lindsay, S., & Grey, S. (2007). A randomised placebo-controlled trial of a self-help Internet-based intervention for test anxiety. *Behaviour Research and Therapy, 45*(3), 483–496. https://doi.org/10.1016/j.brat.2006.04.002.
*Prochaska, J. O. (1971). Symptom and dynamic cues in the implosive treatment of test anxiety. *Journal of Abnormal Psychology, 77*(2), 133. https://doi.org/10.1037/h0030738.
R Core Team (2013). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.
*Rajiah, K., & Saravanan, C. (2014). The effectiveness of psychoeducation and systematic desensitization to reduce test anxiety among first-year pharmacy students. *American Journal of Pharmaceutical Education, 78*(9).
*Reed, M., & Saslow, C. (1980). The effects of relaxation instructions and EMG biofeedback on test anxiety, general anxiety, and locus of control. *Journal of Clinical Psychology, 36*(3), 683–690. https://doi.org/10.1002/1097-4679(198007)36:3<683::aid-jclp2270360313>3.0.co;2-p.
*Reiss, N., Warnecke, I., Tolgou, T., Krampen, D., Luka-Krausgrill, U., & Rohrmann, S. (2017). Effects of cognitive behavioral therapy with relaxation vs. imagery rescripting on test anxiety: A randomized controlled trial. *Journal of Affective Disorders, 208*, 483–489. https://doi.org/10.1016/j.jad.2016.10.039.
Review Manager (RevMan) [Computer program] (Version 5.3.5). (2014). Copenhagen:

The Nordic Cochrane Centre: The Cochrane Collaboration.
*Ricketts, M. S., & Galloway, R. E. (1984). Effects of three different one-hour single-session treatments for test anxiety. *Psychological Reports, 54*(1), 115–120.
*Romano, J. L., & Cabianca, W. A. (1978). EMG biofeedback training versus systematic desensitization for test anxiety reduction. *Journal of Counseling Psychology, 25*(1), 8–13.
*Russell, R. K., & Lent, R. W. (1982). Cue-controlled relaxation and systematic desensitization versus nonspecific factors in treating test anxiety. *Journal of Counseling Psychology, 29*(1), 100–103.
*Russell, R. K., Wise, F., & Stratoudakis, J. P. (1976). Treatment of test anxiety by cue-controlled relaxation and systematic desensitization. *Journal of Counseling Psychology, 23*(6), 563–566. https://doi.org/10.1037//0022-0167.23.6.563.
Rückert, H.-W. (2015). Students' mental health and psychological counselling in Europe. *Mental Health & Prevention, 3*(1), 34–40. https://doi.org/10.1016/j.mhp.2015.04.006.
*Sapp, M. (1996). Three treatments for reducing the worry and emotionality components of test anxiety with undergraduate and graduate college students: Cognitive-behavioral hypnosis, relaxation therapy, and supportive counseling. *Journal of College Student Development, 37*(1), 79–87.
*Saravanan, C., & Kingston, R. (2014). A randomized control study of psychological intervention to reduce anxiety, amotivation and psychological distress among medical students. *Journal of Research in Medical Sciences, 19*(5), 391–397.
Saravanan, C., Kingston, R., & Gin, M. (2014). Is test anxiety a problem among medical students: A cross sectional study on outcome of test anxiety among medical students? *International Journal of Psychological Studies, 6*(3), 24–31.
Schaefer, A., Mattheg, H., Pfitzer, G., & Koehle, K. (2007). Mental health and performance of medical students with high and low test anxiety. *Psychotherapie Psychosomatik Medizinische Psychologie, 57*(7), 289–297. https://doi.org/10.1055/s-2006-951974.
Schulz, K. F., Altman, D. G., Moher, D., & Grp, C. (2010). CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. *BMC Medicine, 8*. https://doi.org/10.1186/1741-7015-8-18.
Seipp, B. (1991). Anxiety and academic performance: A meta-analysis of findings. *Anxiety Research, 4*(1), 27–41. https://doi.org/10.1080/08917779108248762.
Suicide by Children and Young People in England (2016). *National Confidential Inquiry into Suicide and Homicide by People with Mental Illness (NCISH)*. Manchester: University of Manchester.
Temple, J., Salmon, P., Tudur-Smith, C., Huntley, C. D., & Fisher, P. L. (2018). A systematic review of the quality of randomized controlled trials of psychological treatments for emotional distress in breast cancer. *Journal of Psychosomatic Research, 108*, 22–31. https://doi.org/10.1016/j.jpsychores.2018.02.013.
Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48.
von der Embse, N., Jester, D., Roy, D., & Post, J. (2018). Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review. *Journal of Affective Disorders, 227*, 483–493. https://doi.org/10.1016/j.jad.2017.11.048.
Wells, A. (2000). *Emotional disorders and metacognition: Innovative cognitive therapy*. Chichester, UK: Wiley.
Wells, A. (2009). *Metacognitive therapy for anxiety and depression*. New York: Guilford Press.
Wolpe, J. (1958). *Psychotherapy by reciprocal inhibition*. Stanford, CA: Stanford University Press.
Wolpe, J. (1969). Basic principles and practices of behavior therapy in neuroses. *American Journal of Psychiatry, 125*(9), 1242. https://doi.org/10.1176/ajp.125.9.1242.
Zeidner, M. (1998). *Test anxiety: State of the art*. London, UK: Kluwer Academic Publishers.
Zeidner, M., & Matthews, G. (2005). Evaluation anxiety. In A. J. Elliot, & C. S. Dweck (Eds.). *Handbook of competence and motivation*. London, UK: Guildford Press.