Case Study 1: An Evidence-Based Practice Review Report

Theme: School Based Interventions for Learning

The effectiveness of school-based interventions for reducing test anxiety in primary school pupils: a systematic review

Summary

Test anxiety derives from a complex interplay of cognitive, affective-physiological and behavioural components, which transforms a test into a threat (Zeidner & Matthews, 2005). This phenomenon has been researched for over 40 years, yet as more countries adopt attainment-focused education policies, never has this work been more important. In the UK, high-stakes testing at Key stage 2 (Year 6) has led to increasingly high levels of school-related anxiety and stress, disaffection and mental health problems (Hutchings, 2015) and this needs to be addressed. A systematic literature review was conducted to examine the effectiveness of school-based interventions to reduce test anxiety in primary school pupils. Six studies were included and evaluated using a Weight of Evidence framework (Gough, 2007). These studies assessed a range of different interventions to reduce test anxiety: cognitive-behavioural techniques (Aydin & Yerin, 1994), relaxation techniques (Carsley, Heath, & Fajnerova, 2015; Birtürk & Karagün, 2015; Glanz, 1994) and classroom approaches (Putwain & Best, 2012; Yeo, Goh, & Liem, 2016). The review found evidence that school-based interventions significantly reduce test anxiety for pupils in primary schools, especially amongst those with high-test anxiety. Further research must address the limitations highlighted by this review.

Introduction

Like it or loathe it, sitting an exam is a common experience but for some, those who experience 'test anxiety', it can be debilitating (Aydin & Yerin, 1994). Over the past 40 years, researchers have examined the phenomenon of test anxiety and in particular, its link to performance (Hembree, 1988).Today, education systems around the world are more attainment-focused than ever, which leads to rising pressures within home and school environments. Recent educational policy changes in the UK have meant that the 'high-stakes' testing of secondary school pupils has been replicated in the primary sector, with Standard Assessment Tests (SATs) occurring at Year 2 and Year 6. This approach helps to make schools more accountable for performance but teachers and pupils can feel the impact (Hutchings, 2015). Concerns have been raised that pupils as young as age 7 demonstrate symptoms of 'test anxiety' (Connor, 2003) and it is estimated that up to one third of pupils within primary schools experience test anxiety (Lowe, Lee, Witteborg, Pritchard, & Luhr, 2008).

Definitions of test anxiety have changed over time but there is agreement that it is a specific anxiety that develops before an upcoming test. This anxiety derives from a complex interplay of cognitive, affective-physiological and behavioural components, which transform the notion of a test from a benign challenge into a threat (Zeidner & Matthews, 2005). Whilst during testing, most people experience a certain level of anxiety, those suffering from test anxiety may experience heightened levels of self-doubt, rapid heartbeat or sweaty palms and off task behaviours such as fidgeting or staring (Salend, 2011). In children, test anxiety can manifest itself as tearfulness, a

reluctance to go to school or a lack of concentration (Connor, 2003). Pressures from both parents and teachers can heighten these feelings.

Much of the research has focused on performance and has shown that high levels of test anxiety can lead to reductions in test performance compared to those with average anxiety levels (Hembree, 1988). In particular, important or high-stakes tests can lead to even higher test anxiety and reduction in performance (Segool, Carlson, Von Der Embse & Barterian, 2013). Gender, ethnicity and learning disabilities all effect levels of test anxiety (Hembree, 1988; Sena, Lowe, & Lee, 2007).

Test anxiety has also been linked to longer-term emotional issues and Beidel and Turner (1988) found that 60% of highly anxious children in their study also had a diagnosis of generalised anxiety disorder**.** Test anxiety is also considered a mental health condition amongst Singaporean school children (Yeo et al., 2016). In the UK, Childline conducted 9% more counselling sessions about exam anxiety in 2016 compared to the previous year and warned that exam stress can lead to self-harm and suicidal feelings (NSPCC, 2016). The pressure to succeed in school exams brings about a complex combination of emotional, physiological and behavioural reactions, so how can these pressures be reduced for primary school children?

Test anxiety Interventions

Since the 1950s, varieties of test anxiety interventions have been researched including systematic desensitization, rational emotive therapy (to change thoughts) and cognitive and behavioural therapies. More recently, the focus has shifted to a broader generalised anxiety approach and techniques such as Mindfulness, Growth Mindset and Resilience programmes have shown some promise in this area.

Meta-analyses conducted by Hembree (1988) and Ergene (2003) covering 193 studies reported that behavioural-skill-based and cognitive-behavioural interventions produced the highest effect sizes amongst University students. A more recent systematic literature review (Von der Embse, 2013) evaluated interventions (n = 10) between 2000 and 2010 and found promising results for interventions based upon behavioural theory, cognitive theory, cognitive-behavioural theory, academic-skill building and biofeedback. However, only two of the studies investigated reductions in test anxiety in primary/elementary school pupils. These studies showed reduced anxiety in third grade students who were taught relaxation techniques (Larson, Ramahi, Conn, Estes and Ghibelini, 2010) and students with dyslexia reduced their spelling-specific test anxiety by undertaking extra spelling tuition (Faber, 2010).

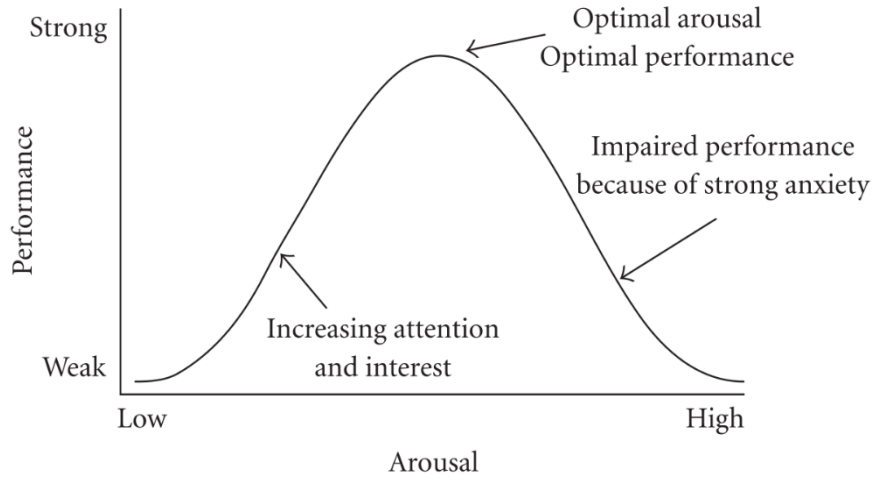
As shown, previous research has mainly investigated test anxiety in secondary and higher education students. There has been relatively little research to examine interventions to reduce test anxiety in primary school pupils.

Psychological theory

Certain levels of anxiety should be expected during testing and this can lead to improvements in performance (Cizek & Burg, 2006). Test anxiety can be explained by the Yerkes-Dodson curve (1908) in Figure 2. Increasing levels of anxiety can lead to enhanced performance initially, such that the rise in adrenaline prior to an exam can be helpful. However, once a particular level of performance has been reached, the levels of anxiety negatively impact upon test performance.

Figure 1

Yerkes-Dodson Curve of optimal performance (Yerkes-Dodson, 1908).



This is also explained by cognitive load, as test anxiety has been shown to increase cognitive demands due to responses to the threat (test). This reduces the capacity of the working memory (Swanson, Hoskyn & Lee, 1999) to respond to the demands of the test, which leads to reduced performance. By being able to reduce the cognitive 'worrying' components of test anxiety, pupils would be better equipped to manage their test performance.

These emotional components of test anxiety can also be affected by the academic self-concept as for many pupils, their level of 'self-worth' is attached to their level of academic achievement (Aydin & Yerin, 1994). Children who have lower perceptions of their academic abilities have been found to be more likely to suffer from test anxiety and have reduced exam performance (Putwain, 2008).

An intervention to reduce test anxiety that helps pupils understand their own physiological responses to threat, educates them in stress-management techniques and reduces negative thought reduction techniques, would be transformative.

Rationale for Review

In order to gain a better understanding of the test anxiety interventions that have been conducted in primary schools, it is important to undertake a systematic literature review. Previous systematic literature reviews (Von der Embse, 2013) have only found limited studies that reviewed interventions to reduce test anxiety in primary-age children. Most of the research in this field has been conducted with university or secondary age students, which whilst useful, is not helpful in determining effective interventions for primary school children.

With high-stakes testing now taking place in primary schools, it is evident that the best place to conduct interventions to reduce test anxiety should also occur in the primary school setting (Weems, Scott, Taylor, Perry, & Triplett, 2010). By identifying pupils at risk of test anxiety at this early age, interventions can help reduce anxiety and better prepare pupils for test taking in the future. Recent UK Government directives, such as 'Healthy schools' (Arthur et al., 2011) have focused on well-being and, given the link between high test anxiety and reduced performance (Hembree, 1988), it is anticipated that schools may look to professionals to offer advice on effective, evidence-based interventions to reduce test anxiety.

Understanding interventions to reduce test anxiety in primary school pupils is important for Educational Psychologists (EPs) as the continued focus on attainment

and 'high-stakes testing' may lead to a greater number of young people experiencing social, emotional and mental health issues which will affect their learning and development. EPs, as professionals working directly with schools, may be asked to recommend interventions to support pupils prior to test-taking, however, as little research currently exists relating to test anxiety in this age group (age 4-11), EPs may not currently be able to recommend appropriate interventions. This review seeks to examine and evaluate a range of test anxiety interventions for primary age children that will enable EPs to determine the most appropriate interventions for their setting. At the individual level, this could be helping to support a young person who is experiencing test anxiety, recommending a research-based, whole class intervention or educating staff on the signs and symptoms of test anxiety. EPs can support interventions by helping the school to identify potential pupils who are at risk of suffering from test anxiety for example pupils with learning needs (Hembree, 1988). Where costs allow, EPs can administer the intervention and assist in the evaluation of outcomes. The current review aims to examine the available evidence to determine whether the school-based interventions address the issues raised by these psychological theories in order to reduce test anxiety in primary school pupils.

Review Question: How effective are school-based interventions for reducing test anxiety in primary school pupils?

Critical Review of the Evidence Base

Literature search

To address the current review question, a literature search was conducted on 20th January 2017. A keyword search examined databases: PsychINFO, ERIC and British Educational Index. Keywords included variations on the phrase 'test anxiety',

variations on the word 'intervention' and as the participants are considered to be children aged 5-11 in primary school (UK), the words 'primary school' and 'elementary school', (which is the equivalent school in other parts of world) were included. These keywords were combined with use of the word AND, to focus the search. Table 1 shows the exact search terms used for PsychINFO and these were amended for each database to retain the same concept.

Table 1

Search terms for PsychINFO search

Outcome		**Participants**		**Intervention**
test* OR		intervent* OR		primary school OR
anxi* OR	AND	program* OR	AND	elementary school OR
stress* OR		reduc*		
exam* OR				
(test anxiety)				

* was used to denote all variations of that word for example 'exam*' could be 'exams' or 'examination' or 'examinations'

A filter was applied to restrict the search to peer reviewed journals, written in English and published after 1990. The search generated 50 results across the three databases. Seven studies were excluded, as they were duplicates. Titles and abstracts were then screened using the inclusion and exclusion criteria (Table 2), which resulted in 12 studies being retained for full text review. A flow diagram of the full search is shown in Figure 1 and a list of excluded studies with rationale is contained in Appendix A. Of these 12, six studies were selected for data synthesis. Table 3 outlines details of each of the included studies.

Table 2

Inclusion and Exclusion criteria

1	Participants	Children aged 5 years -11 years	Children aged <5 years and >11 years	Test anxiety has been identified in children as young as 7 (Connor, 2003) and this is a focus of this review
2	Setting	Interventions conducted in Primary/Elementary schools	Interventions not conducted in Primary/Elementary schools	As the tests/exams take place in school, the interventions should also be conducted in the school setting (Weems et al., 2010).
3	Intervention	Any intervention to reduce general test anxiety prior to the test being taken	Any intervention that does not specifically target test anxiety, targets a specific subject test anxiety or is not conducted prior to a test	The impact of test anxiety interventions are the focus of this review
4	Study design	-Must be a primary study of an intervention to reduce test anxiety -Must include appropriate quantitative measures to assess the effectiveness of the intervention - Must include a treatment group and an appropriate control	- Secondary studies e.g. systematic literature review, meta analyses. - Does not include appropriate quantitative measures to assess the effectiveness of the intervention - No appropriate control - Qualitative studies	-This review will contain empirical studies to ensure originality of findings. -Without appropriate quantitative measures or an appropriate control, it will not be possible to evaluate the effectiveness of the intervention in reducing test anxiety.
5	Publication language	Article published in English	Article published in a language other than English	Translation services are not available. English is the primary language of the reviewer
6	Published date	From 1990 to February 2017	Before 1990 or after February 2017	-By including documents published after 1990, it will allow a wide range of studies to be reviewed. -The review must be completed by 20th February 2017
7	Type of publication	Peer reviewed journal	Non-peer reviewed journal, dissertations, book chapters	Peer-reviewed journals have higher credibility based on their comprehensive assessment process
8	SLR inclusion status	Not included in a previous systematic literature review	Included in a previous systematic literature review	To ensure that the data has not already been critiqued
9	Country	Research completed in an OECD country	Research not completed in an OECD country	Non-OECD Education systems may be very different to OECD Education systems so the research results may not be as applicable

Figure 2

Flow diagram of study selection process

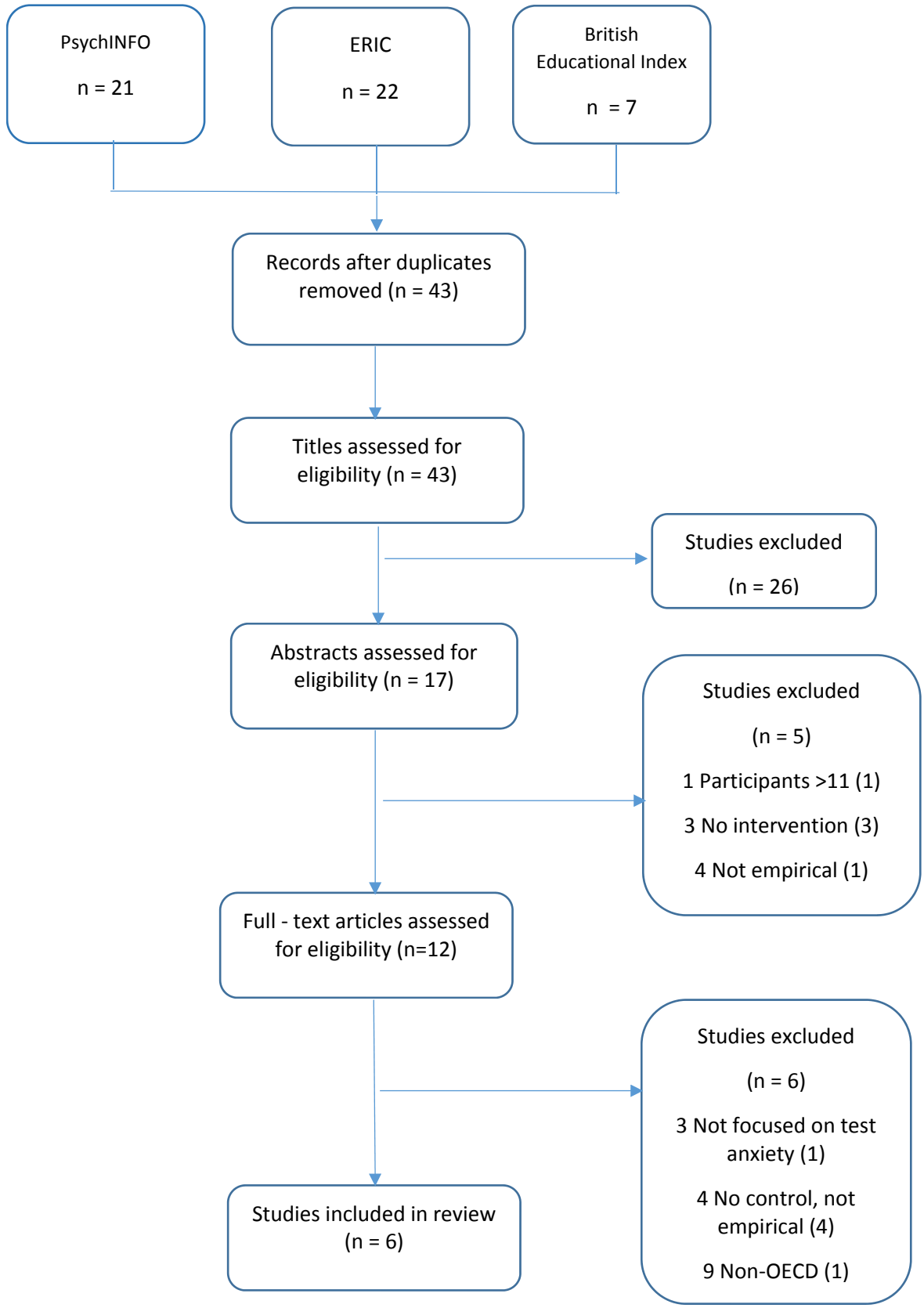


Table 3

Included studies examining school-based interventions to reduce test anxiety

No.	Author	Country	Participants	Study design	Intervention	Measures	Primary Outcomes
1.	Aydin and Yerin (1994)	Turkey	N = 20 5th grade students Age 11-12 55% male High anxiety Middle class	Group-based experimental design Intervention group and control group (no intervention) Random assignment to conditions	Story-based cognitive-behaviour modification comprising continuous story, discussions, breathing activities, elimination of negative thoughts 8 weeks , 1 x 45min session per week	TAI - Turkish version for Total Test Anxiety, Worry and Emotionality. Data collection: Pre and post – 2 weeks before exam. NB: The exam was cancelled and re-arranged so a follow measure was added 3 days before 2nd exam to check for impact	Significant reduction in test anxiety in intervention group compared to control at post and follow up
2.	Beidel et al. (1999)	USA	N= 8 4-6th grade students Mean age 10.9 years 50% female Moderate to severe test anxiety Upper middle class Caucasian Average academic ability	Within participants group design (repeated measures) Control – 6 month period prior to intervention measured by TASC/Physiological tests	11 week Testbusters treatment programme including study habits, SQ3R teaching, test-taking strategies Weekly group meetings plus additional 15-20 minutes study at home	6 months prior to intervention TASC and ADISC administered to determine level of test anxiety Re-administered before start of intervention (as control) PSCSC and physiological tests (pulse rate when completing tasks)	TASC scores decreased significantly over the 11 week intervention $t(df=7) = 5.24$, $p < .001$

No.	Author	Country	Participants	Study design	Intervention	Measures	Primary Outcomes
3.	Carsley et al. (2015)	Canada	N = 52 4-6th grade students Mean age: 10.92 years 53.8% female Middle class 78% English as first language	Group-based experimental Intervention group and comparative control Random assignment to conditions	Mindfulness based colouring activity for groups (15-21 pupils per group) Colouring (structured mandala v unstructured) for 15 minutes prior to spelling test Anxiety induced by telling pupils that their results would be shared with parents	STAIC-S – completed pre colouring and post colouring (prior to test) Spelling test results (WRAT-3)	Significant decrease in test anxiety for both intervention and control groups Test anxiety significantly reduced in Mandala condition. Test anxiety greater reduction in the free colouring condition for boys
4.	Glanz (1994)	USA	N = 28 Fifth grade 100% male Learning disabled students (2-3 years below grade in Reading and 1-2 years below grade in Maths plus concentration and behaviour problems) African-American	Group-based experimental Intervention group and control (no intervention) Stratified random sampling procedure used to divide group into low/middle/high anxiety levels and randomly assigned to 2 classes	Whole class stress-reduction training (self-management of breathing, concentration, Tai Chi) 1 x 30 min instruction per day for 12 weeks Exam in May-Degrees of Reading Power (DRP) test	TASC administered in September, December, March and May (prior to DRP test) LSC DSC	Significant difference in test anxiety between the groups. The anxiety levels of the control group increased throughout the year towards the test, the experimental group increased initially then fell.

No.	Author	Country	Participants	Study design	Intervention	Measures	Primary Outcomes
5.	Putwain and Best (2012)	UK	N =38 Year 3(9) Year 4(10) Year 5(10) Year 6(9) 58% male Mid ability range	Mixed design, one between factor (low v high trait test anxiety) and one within participants factor (high fear vs low fear appeal) 4 intervention groups – order of intervention counterbalanced	Impact of Fear appeals (teachers constantly remind pupils about the importance of exams) on test anxiety and performance Intervention conducted for 2 weeks - One week of maths lessons in one condition followed by Maths test then alternate condition	CTAS completed 4 weeks prior to intervention to assess anxiety levels STAIC and TUFAQ completed after the test	State anxiety higher in high fear appeals condition than low fear appeals High trait test anxious students reported more frequent and threatening fear appeals than low trait test anxious students
6.	Yeo et al. (2016)	Singapore	N = 115 4th grade Age 9-12 (mean age 10.15years) 60% male	Quasi-experimental design Intervention CBT group and control Non-randomised	Whole class CBT intervention (inc breathing, imaginal desensitization and study skills) plus homework Rewards were given for those who completed activities each day 1 x 30 min lesson per week for 4 weeks (prior to exam)	CTAS and CBSC at 3 data collection points: Pre (one month before test)/post (a few days before test)/ 2 months after intervention (prior to next test)	Reductions in test anxiety in intervention group compared to control. Pre-test to follow up reported moderate effect size, d=0.52 No significant change was reported between pre-testing and post-testing for CBT group

ADISC – Anxiety Disorders Interview Schedule for Children, CBSC – Cognitive Behavioural Skills Checklist , CTAS – Childrens Test Anxiety Scale, DSC – Defensiveness Scale for Children, LSC – Lie Scale for Children, PSCSC – Perceived Self-Competence Scale for Children, SQ3R – Survey, Question, Read, Review, Recite (study method), STAIC-S – Stait-Trait Anxiety Inventory for Children State form, TAI - Test anxiety Inventory, TASC – Test Anxiety Scale for Children, TUFAQ - Teachers Use of Fear Appeals Questionnaire

Weighting of studies

The six selected studies were evaluated according to the Weight of Evidence framework (Gough, 2007), which provides a mechanism for evaluating the extent to which each study contributes to addressing the research question. The framework comprises three areas:

- Weight of Evidence A (WoE A): evaluates the methodological quality
- Weight of Evidence B (WoE B): evaluates the methodological relevance
- Weight of Evidence C (WoE C): evaluates the relevance to the review question.

The group-design studies (5/6) were evaluated for WoE A using the APA Task Force Coding Protocol for group design (Kratochwill, 2003). The protocol was amended to evaluate the quality of measurement, comparison group and the statistical analyses. The methodological quality for the single-case design (Beidel, Turner and Taylor-Ferreira, 1999) was evaluated using an appropriate coding protocol (Horner et al., 2005). For each protocol, scores from all WoE A sections were combined and then averaged to form an overall WoE A score. The WoE framework including rating criteria and rationale for protocol amendments is listed in Appendix C. An example of the completed coding protocols is in Appendix D.

The studies were then evaluated for methodological relevance (WoE B) and the focus area of the study (WoE C) using defined rating criteria (Appendix C). Following the scoring, the WoE dimensions were then averaged to provide an overall score, WoE D, which is a measure of the effectiveness of each study in answering the research question.

Table 4 provides an overview of the weight of evidence scores for the six studies in this review. Further information can be found in Appendices C-E.

Table 4

Weight of Evidence ratings

Study	Weight of Evidence A	Weight of Evidence B	Weight of Evidence C	Weight of Evidence D
Aydin and Yerin (1994)	medium 2	medium 2	medium 2	medium 2
Beidel et al.(1999)	medium 2	low 1	medium 2	medium 1.67
Carsley et al. (2015)	low 1.25	high 3	low 1	medium 1.75
Glanz (1994)	medium 1.75	medium 2	medium 2	medium 1.91
Putwain and Best (2012)	medium 2	medium 2	medium 2	medium 2
Yeo et al. (2016)	medium 2	low 1	medium 2	medium 1.67

Note. Where low up to 1.4, medium 1.5 – 2.4 and high 2.5 +

Characteristics of included studies

Key participants

The six selected studies were published between 1994 and 2016 and reported the results of school-based interventions to reduce test anxiety in USA (n=2), Turkey (n=1), UK (n=1), Singapore (n=1) and Canada (n=1). Two hundred and sixty-one pupils participated in the interventions with sample sizes ranging from eight (Beidel et al., 1999) to one hundred and fifteen pupils (Yeo et al., 2016). Demographic information was specified for most of the studies. Gender proportions ranged from 46.2% male (Carsley et al., 2015) to 100% male (Glanz, 1994). Participants ranged from Years 3-7 (Grades 2nd - 6th), with the mean age 10.5 years. Three studies (Aydin & Yerin, 1994; Beidel et al., 1999; Carsley et al., 2015) reported that participants were middle or upper middle class and a further study (Yeo et al., 2016) took place in a private school. Two studies included pupils of average academic ability (Beidel et al., 1999; Putwain & Best, 2012), one study included pupils of mixed ability (Yeo et al., 2015) and one study (Glanz, 1994) focused on students with learning disabilities. As this is a high risk group for test anxiety (Sena et al., 2007), this study received a higher WoE C rating. Two of the studies specifically included pupils with 'high test anxiety' (Aydin & Yerin, 1994; Beidel et al., 1999) whilst Putwain and Best (2012) compared high and low test anxiety pupils. One study (Yeo et al., 2016) conducted post hoc analysis to evaluate the intervention based on levels of anxiety. All received higher ratings on WoE C and WoE A (Horner et al., 2005). One study (Carsley et al., 2015) did not target high anxiety students, so received lower weightings on WoE C.

Each of the studies recruited samples from one primary/elementary school apart from Beidel et al., (1999) who recruited students from several schools to form an intervention group, which improved their WoE C scores. Studies that did not include demographic details were given a lower rating than those that provided more information.

Study design

Four studies used a pre-post test design where all participants undertook testing at baseline and post intervention (Carsley et al., 2015; Glanz, 1994) or baseline, post intervention and follow up (Aydin & Yerin, 1994; Yeo et al., 2016). The studies that included follow-up attracted a higher WoE B rating as this is important for assessing the longer-term impacts of an intervention. Glanz (1994) also collected data at multiple time-points throughout the year (September, December, March and May). These studies attracted a higher WoE B rating as including pre and post measures enables better analysis of effectiveness than other ratings. One study (Putwain & Best, 2012) used a mixed design and the lowest rating for WoE B was given to a within-participants design (Beidel et al., 1999). All of the studies included adequate controls of either a control group (Aydin & Yerin, 1994; Carsley et al., 2015; Glanz, 1994; Putwain & Best, 2012; Yeo et al., 2012) or control period (Beidel et al., 1999). However, Beidel et al. (1999) received the lowest control rating in WoE A due to limited control of threats to validity. Three studies (Aydin & Yerin, 1994; Glanz, 1994 and Yeo et al., 2015) gave no intervention to their control groups, whilst Carsley et al. (2015) used a weaker form of the mindfulness intervention as an alternative during their study. This attracted a higher rating on WoE A and B due to improved controls.

Interventions

The interventions included in this study cover three main areas – cognitive-behavioural therapy, relaxation and classroom-based approaches. Two studies (Aydin & Yerin, 1994; Yeo et al., 2016) used cognitive-behaviour therapy techniques. Aydin and Yerin (1994) implemented a CBT story-based intervention over 8 weeks (1 x 45 minute session per week) and Yeo et al., (2016) conducted 4 weeks of cognitive and behavioural training with

pupils (1 x 30 minute session per week). Two studies (Carsley et al., 2015; Glanz, 1994) examined the impact of relaxation techniques. Carsley et al. (2015) examined the effectiveness of a 15-minute mindfulness colouring activity prior to a spelling test and Glanz (1994) explored a 12-week whole class intervention for stress-management prior to a test including breathing, concentration and Tai Chi. Classroom-based approaches were assessed by two studies (Putwain & Best, 2012; Beidel et al., 1999). Putwain and Best (2012) investigated how 'fear appeals' (reminders about the importance of exams from the teacher) affected high and low anxiety pupils over a 2 week period and Beidel et al. (1999) examined the impact of a study-skills programme 'Testbusters' on pupils over 11 weeks.

Key to the intervention was the requirement for an upcoming test of high importance. Only three studies (Aydin & Yerin, 1994; Glanz, 1994; Yeo et al., 2016) conducted interventions in the run up to an important test. One study (Beidel et al., 1999) conducted testing in an appropriate year for a high-stakes test, whilst two studies (Carsley et al., 2015; Putwain & Best, 2012) created tests for this intervention. The WoE C ratings reflect this.

Measures

All of the studies used an appropriate measure of test anxiety. Two studies (Putwain & Best, 2012; Carsley et al., 2012) used the State-Trait Anxiety Inventory for children (STAIC: Spielberger, Gorsuch, Lushene, Vagg, & Jacobs, 1983) and Aydin and Yerin (1994), used an older version of this test, Test Anxiety Index (TAI). Two studies (Glanz, 1994; Beidel et al., 1999) used the Test Anxiety Scale for Children (TASC: Sarason, Davidson, Whitehall & Waite, 1958). Two studies (Putwain & Best, 2012; Yeo et al., 2016) used the Childrens Test Anxiety Scale (CTAS: Wren & Benson, 2004). Putwain and Best

(2012) used the CTAS as a baseline measure to collect data on anxiety levels and then used STAIC in the intervention. All studies reported reliability coefficients for these measures, however they also only used a single measure to collect comparison data and this has been reflected in their WoE A scores. Beidel et al. (1999) also collected data using the Anxiety Disorders Interview Schedule for Children and collected physiological measures to capture pulse rates at heightened times of stress levels as a baseline. This is reflected in the WoE A baseline score. The transcripts of these interviews were not made available, so they cannot be reported in this review.

Outcomes

All of the studies reported statistically significant reductions in test anxiety. Most studies (Aydin & Yerin, 1994; Carsley et al., 2015; Glanz, 1994; Yeo et al., 2016) reported the effect size, however, to ensure consistency, effect sizes were re-calculated using the Pretest-Posttest Control Group calculation (Morris, 2007). This calculated the standardised mean difference by dividing the mean pre-post difference in intervention and control groups by the pooled standard deviation of the pre-test scores. As Beidel et al. (1999), did not include a control group, the within-person change (Becker, 1988) was calculated as the effect size. Here, the standardised mean difference was calculated by dividing the pre-post mean difference by the standard deviation of the pre-test score. Where it was not possible to calculate the effect size due to insufficient data (Putwain & Best, 2012), the author reported the cited effect size, partial eta squared (η_p^2). Cohen's *d* (1988) effect size descriptors have been used to describe the standardised mean difference, where small, d = .2, medium, d = .5 and large, d = .8. For partial eta squared (η_p^2), the effect size descriptors (Cohen, 1998) are small = 0.01, medium = 0.06, large = 0.14.

A large effect was reported by Glanz (1994) for the stress-reduction training on test anxiety (d = -1.72) and an even larger effect was reported by Beidel et al. (1999) for the impact of Testbusters treatment programme on test anxiety (d = -2.37). However, it should be noted that this was a within-person change, so may not be as robust as other control measures. A medium effect of test anxiety was reported by one of the CBT interventions (Aydin & Yerin, 1994) compared to the control group. Yeo et al. (2016) reported an effect size of <0.2 which is negligible. Effect sizes increased for both studies between post-intervention and follow-up (d = -0.63 v -0.59 and d= -0.41 v -0.14) which may indicate a lag effect for CBT interventions. Carsley et al. (2015) showed a significant decrease in test anxiety in both intervention and control groups but the effect size was negligible (d = -0.07). This may have been affected by the time lag between the pre and post measures, which was only 15 minutes. Putwain and Best (2012) showed that state anxiety was higher in the high fear appeals group compared to the low fear appeals group, indicated by the significant main effect reported for perceived frequency of the fear appeal ($p > 0.001$, $\eta_p^2 = .61$) and perceived threat ($p > 0.001$, $\eta_p^2 = .63$). The effect sizes for each study are presented in Table 5.

Table 5

Effect sizes and Descriptors for Statistically Significant Figures

Type of Study	Author	Participants	Measure	Comparison	Effect size	Descriptor[a]	WoE D
CBT	Aydin and Yerin. (1994)	20	TAI	Intervention vs control group (pre vs post intervention)	*d* = -0.59	medium	medium 2
			TAI	Intervention vs control group (pre vs 2 weeks follow up)	*d* = -0.63	medium	
CBT	Yeo et al. (2016)	115	CTAS	Intervention vs control group (pre vs post intervention)	*d* = -0.14	negligible	medium 1.67
			CTAS	Intervention vs control group (pre vs 2 months follow up)	*d* = -0.41	small	
Relaxation (Mindfulness)	Carsley et al. (2015)	52	STAIC-S	Intervention vs control group (pre vs post intervention)[b]	*d* = -0.07	negligible	medium 1.75
Relaxation (Breathing, Tai Chi)	Glanz (1994)	28	TASC	Intervention vs control group (pre vs post intervention)[c]	*d* = -1.72	large	medium 1.91
Class Approaches (Pupil)	Beidel et al. (1999)	8	TASC	Repeated measures (pre vs post intervention)	*d* = -2.37[d]	large	medium 1.67
Class Approaches (Teacher)	Putwain and Best (2012)	38	CTAS STAI	Within participants (Low v high fear appeals) Perceived frequency Perceived threat	η_p^2 = 0.61 η_p^2 = 0.63	large	medium 2

[a]Effect size descriptors for *d* (Cohen,1988) are small, d = .2, medium, d = .5 and large, d = .8 and descriptors for η_p^2 (Cohen, 1998) are small = 0.01, medium = 0.06, large = 0.14.

[b]The time period between pre and post STAIC-S measure was 15 minutes [c]The time period between pre and post TASC measure was 9 months

[d]Denotes within-person change (Becker, 1998)

Conclusion and Recommendations

The purpose of this review was to determine the effectiveness of school-based interventions in reducing test anxiety amongst primary school pupils. This review evaluated six studies and all showed significant reductions in self-reported test anxiety. The studies with the highest weight of evidence (Aydin & Yerin, 1994; Putwain & Best, 2012) had medium and large effect sizes. The majority of studies included analysis of pupils in 'risk' groups for example those with above average test anxiety (Putwain & Best, 2012; Aydin et al., 1994; Beidel et al., 1999; Yeo et al., 1994) or those with learning disabilities (Glanz, 1994). This evidence seems to suggest that interventions to reduce test anxiety can be effective for pupils with high test anxiety in primary schools. Caveats must be noted that due to small sample sizes, all of the studies apart from Glanz (1994) were underpowered.

Interventions to address exam anxiety consisted of cognitive-behavioural therapies, relaxation and classroom-based approaches. The CBT interventions (Aydin & Yerin, 1994; Yeo et al., 2016) had medium effect sizes (d = -0.63 and d=0.41 respectively) when pre vs follow up measures were compared. However, a lower effect size was reported at pre vs post intervention, potentially indicating a lag effect. The largest effect size was calculated for the Testbusters programme (Beidel et al., 1999) however this was a within participants design, so should be treated with caution due to validity issues. When comparing study duration, the effect sizes were largest for those of longest duration 11 weeks (Beidel et al., 1999) and 12 weeks (Glanz, 1994) which may indicate that test anxiety reduction needs to have longevity to produce effective changes in anxiety levels. The studies that conducted interventions in the run up to important tests (Aydin & Yerin, 1994; Glanz, 1994; Yeo et al., 2016) showed medium to large effect sizes. This could also reflect the length of the

studies that ran from 4 weeks (Yeo et al., 2016) to 8 weeks (Aydin & Yerin, 1994) and 12 weeks (Glanz, 1994).

Areas for Further Research

The limited number of studies found to examine the effects of test anxiety on primary school pupils demonstrates that there is a need for further research in this area. Whilst the studies examined here have found significant results, they all contain small sample sizes, so should be treated with caution. This review has demonstrated that interventions that target primary school pupils who are most anxious, show the greatest benefit. However, the majority of studies did not examine the longer-term impact of these interventions and future research should address this. The focus on developing Healthy schools in the UK and resilience training for primary school pupils is a positive step forward but research should be conducted to investigate the impact of resilience training on test anxiety to ensure that this initiative is tackling this phenomenon. Educational Psychologists have a role to play in educating primary settings about i) the risks of test anxiety, ii) the measures to identify pupils at risk of developing this debilitating anxiety plus iii) evidence-based recommendations for intervention.

In terms of limitations of this review, firstly, some effective interventions may have been missed due to the scope of the inclusion and exclusion criteria. It is noted that there are recent dissertations and unpublished works that may enhance this review but these were not included. Secondly, with just one reviewer, there could be inherent researcher bias. Thirdly, the sample of six studies is small and as the majority were conducted outside of the UK, it is difficult to generalise any findings to the UK. However, all of the countries included in the study operate education systems that include high-stakes testing for

primary school pupils, so the findings from other countries should be taken into consideration. Finally, the use of self-report measures of test anxiety introduces further inherent bias into measurement. It is suggested that future reviews address these limitations.

Given the increasing demands on primary school pupils and in particular the focus on high stakes testing at a younger age, attention must be drawn to the potential long term impact of stress levels and its effect on mental health. By targeting interventions to ensure pupils are able to cope with the demands of test anxiety within primary school, this better enables young people to deal with high-stakes testing in the future, which may offer substantial long-term health benefits.

References and Appendices

Arthur, S., Barnard, A. M., Day, N., Ferguson, C., Gilby, N., Hussey, D., Morrell, G., & Purdon, S. (2011). Evaluation of the National Healthy Schools Programme Interim Report. *National Centre for Social Research*, (May), 1–102.

Aydin, G., & Yerin, O. (1994). The effect of a story-based cognitive behavior modification procedure on reducing children's test anxiety before and after cancellation of an important examination. *International Journal for the Advancement of Counselling*, *17*(2), 149–161.

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology*, *41*, 257–278.

Beidel, D. C., & Turner, S. M. (1988). Comorbidity of test anxiety and other anxiety disorders in children. *Journal of Abnormal Child Psychology*, *16*(3), 275–87.

Beidel, D., Turner, S., & Taylor-Ferreira, J.C., (1999). Teaching Study Skills and Test-Taking Strategies to Elementary School Students, *23*(4), 630–646.

Birtürk, A., & Karagün, E. (2015). The effect of recreational activities on the elimination of state-trait anxiety of the students who will take the SBS placement test. *Educational research and reviews*, *10*(7), 894–900.

Carsley, D., Heath, N. L., & Fajnerova, S. (2015). Effectiveness of a Classroom Mindfulness Coloring Activity for Test Anxiety in Children Effectiveness of a Classroom Mindfulness Coloring Activity for Test Anxiety in Children. *Journal of Applied School Psychology*, *31*(3), 239–255.

Cizek, G., & Burg, S. (2006). Addressing test anxiety in a high-stakes environment: strategies for classrooms and schools. Thousand Oaks, CA: Corwin Press.

Cohen, J. (1988). Statistical power analysis for the behavioural sciences (2nd ed.). Hillside, NJ: Erlbaum.

Connor, M. J. (2003). Pupil stress and standard assessment tasks (SATs) An update. *Emotional and Behavioural Difficulties*, *8*(2), 101–107.

Ergene, T. (2003). Effective Interventions on Test Anxiety Reduction. *School Psychology International*, *24*(3),313.

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.

Glanz, J. (1994). Effects of stress reduction strategies on reducing test-anxiety among learning-disabled students. *Journal of Instructional Psychology*, *21*(4).

Gough, D. (2007). Weight of evidence: a framework for the appraisal of the quality and relevance of evidence. *Applied and Practice-Based Research. Special Edition of Research Papers in Education*, *22*(2), 213–228.

Hembree, R. (1988). Correlates, causes, effects and treatment of test anxiety. Review of Educational Research, 58(1), 47-77.

Horner, R.H., Carr, E.G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The Use of Single-Subject Research to Identify Evidence-Based Practice in Special Education.

Hutchings, M. (2015). Exam Factories? The Impact of Accountability Measures on Children and Young People, 76.

Kratochwill, T. R. (2003). Task force on evidence-based practice interventions in school psychology.Larson, H., Ramahi, M.,Conn, S.,Estes, L., & Ghibellini, A. (2010). Reducing test anxiety among third grade students through the implementation of relaxation techniques. Journal of School Counselling, 8, 1-19.

Lowe, P.A., Lee, S.W., Witteborg, K.M., Pritchard, K.W, Luhr, M.E., et al. (2008).The Test Anxiety Inventory for Children and Adolescents (TAICA): Examination of the psychometric properties of a new multidimensional measure of test anxiety among elementary and secondary school students. *Journal of Psycheducational Assessment*, 26, 215-230.

Morris, S. B. (2007). Estimating Effect Sizes from Pretest-Posttest-Control Group Designs. *Organizational Research Methods*, *11*(2), 364–386.

NSPCC (2016). Childline reveals increase exam stress counselling last year *nspcc.org.uk* Retrieved from https://www.nspcc.org.uk/fighting-for-childhood/news-opinion/childline-reveals-increase-exam-stress-counselling-last-year/.

Putwain, D. (2008). Examination stress and test anxiety. *The Psychologist*, *21*(12), 1026–1029.

Putwain, D. W., & Best, N. (2012). Do highly test anxious students respond differentially to fear appeals made prior to a test?, *88*(88), 1–11.

Salend, S.J. (2011). Addressing test anxiety. *Teaching Exceptional Children, 44 (2), p. 61.*

Sarason, S. B., Davidson, K., Lighthall, F., & Waite, R. (1958). A test anxiety scale for children. *Child Development*, Mar; 29(1):105-13.

Segool, N. K., Carlson, J. S., Goforth, A. N., von der Embse, N., & Barterian, J. A. (2013). Heightened test anxiety among young children: Elementary school students' anxious responses to high-stakes testing. *Psychology in the Schools*, *50*(5), 489–499.

Sena, J.,D. W., Lowe, P. A., & Lee, S. W. (2007). Significant predictors of test anxiety among students with and without learning disabilities. *Journal of Learning Disabilities, 40*(4), 360-76

Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R., & Jacobs, G. A. (1983). *Manual for the State-Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Swanson, H. L., Hoskyn, M., & Lee, C. (1999). Interventions for students with learning disabilities. New York NY: Guilford Press.

Von der Embse, N., Barterian, J. and Segool, N. (2013), Test Anxiety Interventions for Children and Adolescents: A Systematic Review of Treatment Studies from 2000–2010. Psychol. Schs., 50: 57–71.

Weems, C. F., Scott, B. G., Taylor, L. K., Perry, A. M., & Triplett, V. (2010). Test Anxiety Prevention and Intervention Programs in Schools : Program Development and Rationale, 62–71.

Wren, D. G., & Benson, J. (2004). Measuring test anxiety in children: Scale development and internal construct validation. Anxiety, Stress, and Coping: An International Journal, 17, 227-240.

Yeo, L. S., Goh, V. G., & Liem, G. A. D. (2016). School-Based Intervention for Test Anxiety. *Child & Youth Care Forum*, 1–17.

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. Journal of Comparative and Neurological Psychology, 18, 459-482.

Zeidner, M., & Mathews, G. (2005). Evaluation anxiety. In A.J. Elliot & C.S. Dweck (Eds.) Handbook of competence and motivation. London: Guilford Press.

Appendix A: Excluded Studies

Excluded Studies

Excluded at abstract and full review stages	Code and reason for exclusion
Bhadwal, S. C., & Panda, P. K. (1992). The composite effect of curricular programme on the test anxiety of rural primary school students: a one-year study. *Educational review*, *44*(2), 205–220.	9 Non-OECD country
Birtürk, A., & Karagün, E. (2015). The effect of recreational activities on the elimination of state-trait anxiety of the students who will take the SBS placement test. *Educational research and reviews*, *10*(7), 894–900.	1 Participants > age 11
Buck, R. (2016). Anxiety manipulation in school-based research *Psychology of Education review*	1 Participants > age 11
Cheek, J. R., Bradley, l. J., Reynolds, J., & Coy, D. (2017). An intervention for helping elementary students reduce test anxiety, *Professional School Counseling 6*(2), 162–164.	4. Review of an empirical study
Hobson, S. (1996) Test anxiety: Rain or shine! *Elementary school guidance & counselling, 30*(4), 316-318. Retrieved from http://www.jstor.org/stable/42871234	3 No intervention
Kass, R. G., & Fish, J. M. (1991). Positive reframing and the test performance of test anxious children. *Psychology in the schools*, *28*(1), 43–52.	4 No appropriate control group
Mavilidi M.F., Hoogerheide V., Paas F. (2014), A quick and easy strategy to reduce test anxiety and enhance test performance, Applied Cognitive Psychology, 28, pages 720–726	3 (prior to test not in test)
Nicaise, M. (1995). Treating test anxiety. A review of three approaches. *Teacher education and practice*, *11*(1), 65-81.	4 Not empirical
Putwain, D. W., & Best, N. (2011). Fear appeals in the primary classroom: effects on test anxiety and test grade. *Learning and individual differences*, *21*(5), 580–584.	4 Test anxiety not appropriately measured as a whole construct
Tok, S. (2013). Effects of the know-want-learn strategy on students' mathematics achievement, anxiety and metacognitive skills, *Metacognition and Learning*, 8(2), 193–212.	3 Specific maths test anxiety
Webb, l., Carey, J., Villares, E.(2014) Results of a randomized controlled trial of student success skills, *Society for research on educational effectiveness*. 2014 24 pp.	3 Test anxiety reduction not primary aim

Appendix B: Included studies

Included studies

	Full references of included studies
1.	Aydin, G., & Yerin, O. (1994). The effect of a story-based cognitive behavior modification procedure on reducing children's test anxiety before and after cancellation of an important examination. *International Journal for the Advancement of Counselling*, *17*(2), 149–161.
2.	Beidel, D., Turner, S., & Taylor-Ferreira, J. (1999). Teaching Study Skills and Test-Taking Strategies to Elementary School Students, *Behaviour Modification*. *23*(4), 630–646.
3.	Carsley, D., Heath, N. L., & Fajnerova, S. (2015). Effectiveness of a Classroom Mindfulness Coloring Activity for Test Anxiety in Children. *Journal of Applied School Psychology*, *31*(3), 239–255.
4.	Glanz, J. (1994). Effects of stress reduction strategies on reducing test-anxiety among learning-disabled students. *Journal of Instructional Psychology*, *21*(4).
5.	Putwain, D. W., & Best, N. (2012). Do highly test anxious students respond differentially to fear appeals made prior to a test? *88*(88), 1–11.
6.	Yeo, L. S., Goh, V. G., & Liem, G. A. D. (2016). School-Based Intervention for Test Anxiety. *Child & Youth Care Forum*, 1–17.

Appendix C: Weighting of Evidence framework

Weight of Evidence A (WoE A)

WoE A is a non-review specific judgement (Gough, 2007) that evaluates the quality of the study in general terms. The APA Task Force Review Coding Protocol (Kratochwill, 2003) for group interventions was used to evaluate studies 1, 3,4,5,6 in this review. The Protocol was amended to reflect the studies in this review, which focused on assessments of Measurement (II A), Comparison Groups (II B) and Appropriate Statistical Analyses (II C) and Follow Up Assessment (I). The rationale for these amendments is below:

Amendments to WoE A Coding Protocol

Sections removed	Rationale
I B7- B8	Only quantitative studies are included in the review
II C	This protocol was only used to review the methodological relevance of the study, outcomes are evaluated separately
II D	Only anxiety reduction interventions were included, so this section is not appropriate for this review
E	This review only addresses one component so does not need to be included
F	The studies were all implemented by the researcher so there is no issue of fidelity
G	There was no replication of studies
H	Only school-based interventions were considered as part of this review

Specific criteria for assessing each of the measures below was provided by Kratochwill (2003). A study must meet all the criteria in each weighting to receive that rating

Weighting and criteria for Measurement (II A)

Weighting	Criteria
Strong evidence (3)	- The study used measures that produced reliable scores of at least .7 for the majority of primary outcomes - Data was collected using multiple methods - Data was collected from multiple sources - Validity of measures is reported
Promising evidence (2)	- The study used measures that produced reliable scores of at least .6 for the majority of primary outcomes - Data was collected using multiple methods and/or collected from multiple sources - Validity of measures is reported
Weak Evidence (1)	- The study used measures that produced reliable scores of at least .5 for the majority of primary outcomes - Only one data source or method was used - Validity of measures is not reported
No Evidence (0)	- There is not enough evidence to rate this measure

Weighting and criteria for Comparison Group (II B)

Weighting	Criteria
Strong evidence (3)	- The study includes an 'active' control group - Confidence rating in the type of comparison group is strong - Change agents are counter balanced (where applicable) - There is evidence of group equivalence by random assignment - There is equivalent mortality and low attrition at post and follow up
Promising evidence (2)	- The study includes a control group - Confidence rating in the type of comparison group is moderate - Change agents are not counter balanced (where applicable) - There is evidence of group equivalence through random assignment or post hoc tests - There is equivalent mortality and low attrition at post
Weak Evidence (1)	- The study includes a control group - Confidence rating in the type of comparison group is low - There is no evidence of random assignment - Equivalent mortality is not referenced
No Evidence (0)	- There is not enough evidence to rate this measure

Weighting and criteria for Appropriate Statistical Analysis (II C)

Weighting	Criteria
High (3)	- The study conducted appropriate statistical analysis including an appropriate unit of analysis - Familywise error rate was controlled where appropriate - There was a sufficiently large sample size - Effect sizes were reported for all primary outcomes or there was enough information to calculate effect sizes
Medium (2)	Demonstrated at least 2 of these: - The study conducted appropriate statistical analysis including an appropriate unit of analysis - Familywise error rate was controlled where appropriate - There was a sufficiently large sample size - Effect sizes were included for all primary outcomes or there was enough information to calculate some effect sizes
Low (1)	Demonstrated at least 1 of these: - The study conducted appropriate statistical analysis including an appropriate unit of analysis - Familywise error rate was controlled where appropriate - There was a sufficiently large sample size - Effect sizes were not included for all primary outcomes or there was insufficient information to calculate effect sizes
No Evidence (0)	- There is not enough evidence to rate this measure

Note. Sufficiently large sample size was determined by power calculation using the calculated effect size (*d* or $\eta_p^{2)}$, an alpha value of .05 and power of 0.80. In one study, the power analysis cited in the study was used.

Weighting and criteria for Follow up assessment (I)

Weighting	Criteria
High (3)	- Evidence of follow up data collection at an appropriate time lag - Follow up data was collected from at least 80% original participants - Consistent measures are used to collect data
Medium (2)	Evidence of follow up data collection after intervention - Follow up data was collected from at least 50% original participants - Consistent measures are used to collect data
Low (1)	- Evidence of follow up data collection - Follow up data was collected from at least 20% original participants - Relevant measures are used to collect data
No Evidence (0)	- There is not enough evidence to rate this measure

WoE A overall scores

The scores for each dimension were combined and then averaged to form the overall WoE A score.

Overall weighting for WoE A (Group designs)

Study	Measurement	Comparison Group	Statistical Analysis	Follow Up	Overall WoE A
Aydin and Yerin (1994)	1	2	2	3	medium 2
Carsley et al. (2015)	1	2	2	0	low 1.25
Glanz (1994)	2	2	3	0	medium 1.75
Putwain and Best (2012)	2	3	3	0	medium 2
Yeo et al. (2016)	2	1	2	3	medium 2

Note. Where weak = < 1.4, low = < 1.4, medium =1.5 – 2.4 and high = > 2.

For the single-case design (Beidel et al., 1999) the Horner et al. (2005) coding protocol was used. The following table states the level of rating required for each of the criteria within this protocol.

Rating	Criteria
3 (all criteria met) 2 (2/3 criteria met) 1 (1/3 criteria met) 0 (not enough information to rate)	*Description of Participants and Settings* - Participants are described with sufficient detail to allow others to select individuals with similar characteristics (e.g. age, gender, disability, diagnosis). - The process for selecting participants is described with replicable precision. - Critical features of the physical setting are described with sufficient precision to allow replication.
3 (all criteria met) 2 (3-4/5 criteria met) 1 (1-2/5 criteria met) 0 (not enough information to rate)	*Dependant Variable* - Dependant variables are described with operational precision. - Each dependant variable is measured with a procedure that generates a quantifiable index. - Measurement of the dependant variable is valid and described with replicable precision. - Dependant variables are measured repeatedly over time. - Data are collected on the reliability or interobserver agreement associated with each dependant variable, and IOA levels meet minimal standards (e.g. IOA =80%; Kappa = 60%
3 (all criteria met) 2 (2/3 criteria met) 1 (1/3 criteria met) 0 (not enough information to rate)	*Independent Variable* - Independent variable is described with replicable precision. - Independent variable is systematically manipulated and under the control of the experimenter. - Overt measurement of the fidelity of implementation for the independent variable is highly desirable.
3 (all criteria met) 2 (both criteria almost met) 1 (1/2 criteria met) 0 (not enough information to rate)	*Baseline* - The majority of single-subject research studies will include a baseline phase that provides repeated measurements of a dependent variable and establishes a pattern of responding that can be used to predict the pattern of future performance if introduction or manipulation of the independent variable did not occur. - Baseline conditions are described with replicable precision
3 (all criteria met) 2 (2/3 criteria met) 1 (1/3 criteria met) 0 (not enough information to rate)	*Experimental control/internal validity* - The design provides at least three demonstrations of experimental effect at three different points in time. - The design controls for common threats to internal validity (e.g. permits elimination of rival hypotheses). - The results document a pattern that demonstrates experimental control.

3 (all criteria met) 2 (both criteria almost met) 1 (1/2 criteria met) 0 (not enough information to rate)	*External/validity* - Experimental effects are replicated across participants, settings, or materials to establish external validity. - Selection and attribution biases (e.g., the selection of only certain participants, or the publication of only successful examples) are minimized.
3 (all criteria met) 2 (3/4 criteria met) 1 (1-2/4 criteria met) 0 (not enough information to rate)	*Social Validity* - The dependent variable is socially important. - The magnitude of change in the dependent variable resulting from the intervention is socially important. - Implementation of the independent variable is practical and cost effective. - Social validity is enhanced by implementation of the independent variable over extended time periods, by typical intervention agents, in typical physical and social contexts

Overall weighting for WoE A (Single case design)

Criteria	Beidel et al. (1999)
Participants and setting	medium 2
Independent variable	high 3
Dependent variables	medium 2
Baseline	high 3
Experimental control/ internal validity	weak 0
External validity	medium 2
Social validity	medium 2
Total WoE A	medium 2

Note. Where weak = < 1.4, low = < 1.4, medium =1.5 – 2.4 and high = > 2.5

Weight of Evidence B (WoE B)

WoE B is a review question-specific judgement of research design (Gough, 2007) that gives a score for methodological relevance of the study design to the specific research question. The criteria is based upon evidence hierarchies (Guyatt et al., 2008) which suggests that certain types of studies are more appropriate for addressing particular research questions than others. A study must meet all the criteria in each weighting to receive that rating.

WoE B weighting criteria

Weighting	Criteria
High (3)	- The study includes an 'active' control group - Participants are randomly assigned to the conditions - The groups must be treated equally - The study must report pre and post scores of the primary outcome measure for both groups
Medium (2)	- The study includes a control group - Participants are randomly assigned to the conditions - The study must report appropriate measures to assess the effectiveness of the intervention
Low (1)	- The study design incorporates a control - Participants are not randomly assigned to the conditions - The study must report appropriate measures to assess the effectiveness of the intervention

Rationale for WoE B criteria:

- In order to assess efficacy of the intervention it is important that the treatment group is compared to an 'active' control group, such that any researcher effects or nonspecific effects of conducting the research are included in the study. Studies that include 'no treatment' control groups will have a lower weighting as they will potentially over-estimate the effects of the intervention.
- Participants should be randomly assigned to each condition to ensure that the groups are equivalent and there is no selection bias
- The groups should be treated equally so that the only variable is the intervention , maximising the internal validity
- To establish whether the intervention was effective, pre and post scores of the primary outcome measure should be include for both the treatment and control groups.

Overall weighting for WoE B

Author	Weight of Evidence B
Aydin and Yerin (1994)	medium 2
Beidel et al. (1999)	low 1
Carsley et al. (2015)	high 3
Glanz (1994)	medium 2
Putwain and Best (2012)	medium 2
Yeo et al. (2016)	low 1

Note. Where low = < 1.4, medium =1.5 – 2.4 and high = > 2.5

Weight of Evidence C (WoE C)

WoE C is a review question –specific judgement of evidence focus (Gough, 2007) that gives a score for the relevance of the study to the specific research question. A study must meet all the criteria in each weighting to receive that rating.

WoE C weighting criteria

Weighting	Criteria
High (3)	- Test anxiety is the primary outcome measure for the study - The intervention is conducted prior to an important test/exam - The intervention is documented in sufficient detail to ensure replication - The study includes samples from high risk populations - Participants are included from more than one primary/elementary school - More than one method of data collection is included
Medium (2)	- Test anxiety is one of the key outcome measure for the study - The intervention is conducted in the same year as an important test/exam - The intervention is explained clearly - The study may include high risk groups and the key sample characteristics - Participants are included from at least one primary/secondary school - At least one method of data collection is included
Low (1)	- Test anxiety is one of the outcome measures of the study - The intervention is conducted prior to a test - The intervention is explained - The study includes at least one characteristic of the sample One method of data collection is included

Rationale for WoE C criteria:

- The research question refers to test anxiety therefore the main measure of the study should be test anxiety
- The intervention should be conducted prior to an important exam as this will naturally induce feelings of test anxiety and therefore the effectiveness of the intervention can be monitored. Whilst it would not be ethical to alter test conditions prior to a very important exam, the test/exam should be important enough to induce a test anxiety.
- To ensure replication, the test anxiety intervention should be described in detail otherwise researchers will not be able to repeat the intervention
- High risk populations (Hembree,1988) are at greater risk of test anxiety and therefore by focusing on these groups, it maximises the chances of the intervention being useful in terms of being able to guide both theory and practice
- To ensure there is no selection bias, participants should be recruited from more than one primary/ elementary school
- To ensure reliability of data, more than one method of data collection should be used

Overall weighting for WoE C

Author	Weight of Evidence C
Aydin and Yerin (1994)	medium 2
Beidel et al. (1999)	medium 2
Carsley et al. (2015)	low 1
Glanz (1994)	medium 2
Putwain and Best (2012)	medium 2
Yeo et al. (2016)	medium 2

Note. Where low = < 1.4, medium =1.5 – 2.4 and high = > 2.5

Weight of Evidence D (WoE D)

WoE D gives the score that indicates the extent to which study meets the requirements of the review question. As it is considered that the execution of the study (WoE A), the methodological relevance (WoE B) and focus of the study (WoE C) are all equally important in addressing the research question, the scores for each dimension and combined and then averaged to form the WoE D score.

Overall weight of evidence ratings

Study	Weight of Evidence A	Weight of Evidence B	Weight of Evidence C	Weight of Evidence D
Aydin and Yerin (1994)	medium 2	medium 2	medium 2	medium 2
Beidel et al. (1999)	medium 2	low 1	medium 2	medium 1.67
Carsley et al. (2015)	low 1.25	high 3	low 1	medium 1.75
Glanz (1994)	medium 1.75	medium 2	medium 2	medium 1.91
Putwain and Best (2012)	medium 2	medium 2	medium 2	medium 2
Yeo et al. (2016)	medium 2	low 1	medium 2	medium 1.67

Note. Where low = < 1.4, medium =1.5 – 2.4 and high = > 2.5

Appendix D: Example of completed Coding Protocol

[Adapted from Task Force on Evidence-Based Interventions in School Psychology, American Psychology Association, Kratochwill, T.R. (2003)]

Coding Protocol

Name of Coder:_____ Date:_____

Full Study Reference in proper format: Aydin, G., & Yerin, O. (1994). The effect of a story-based cognitive behavior modification procedure on reducing children's test anxiety before and after cancellation of an important examination. *International Journal for the Advancement of Counselling*, *17*(2), 149–161.

Intervention Name (description of study): Story-based cognitive-behavior modification

Study ID Number:_____1_____

☐ Type of Publication:
☐ Book/Monograph
☐ Journal Article
☐ Book Chapter
☐ Other (specify):

1. General Characteristics

A. General Design Characteristics

A1. Random assignment designs (if random assignment design, select one of the following)

☐ Completely randomized design
☐ Randomized block design (between participants, e.g., matched classrooms)
☐ Randomized block design (within participants)
☐ Randomized hierarchical design (nested treatments

A2. Nonrandomized designs (if non-random assignment design, select one of the following)

☐ Nonrandomized design
☐ Nonrandomized block design (between participants)
☐ Nonrandomized block design (within participants)
☐ Nonrandomized hierarchical design
☐ Optional coding for Quasi-experimental designs

A3. Overall confidence of judgment on how participants were assigned (select one of the following)

☐ Very low (little basis)
☐ Low (guess)
☐ Moderate (weak inference)
☐ High (strong inference)
☐ Very high (explicitly stated)
☐ N/A
☐ Unknown/unable to code

B. Participants

Total size of sample (start of study): _20_____

Intervention group sample size:___10____

Control group sample size:___10_____

C. Type of Program

☐ Universal prevention program
☐ Selective prevention program
☐ Targeted prevention program
☐ Intervention/Treatment
☐ Unknown

D. Stage of Program

☐ Model/demonstration programs
☐ Early stage programs
☐ Established/institutionalized programs
☐ Unknown

E. Concurrent or Historical Intervention Exposure

☐ Current exposure
☐ Prior exposure
☐ Unknown

2. Key Features for Coding Studies and Rating Level of Evidence/Support

(Rating Scale: **3= Strong Evidence, 2=Promising Evidence, 1=Weak Evidence, 0=No Evidence)**

A. Measurement (Estimating the quality of the measures used to establish effects)

A1 The use of the outcome measures produce reliable scores for the majority of the primary outcomes

☐Yes
☐No
☐Unknown/unable to code

A2 Multi-method (at least two assessment methods used)
☐ Yes
☐ No
☐ N/A
☐ Unknown/unable to code

A3 Multi-source (at least two sources used self-reports, teachers etc.)
☐ Yes
☐ No
☐ N/A
☐ Unknown/unable to code

A4 Validity of measures reported (well-known or standardized or norm-referenced are considered good, consider any cultural considerations)

☐ Yes validated with specific target group(Stated normed for age group but cannot find norm stats)
☐ In part, validated for general population only (TAI-T 0.96)
☐ No
☐ Unknown/unable to code

Overall Rating for measurement__1__

3= Strong Evidence 2=Promising Evidence 1=Weak Evidence 0=No Evidence

B. Comparison Group

B1 Type of Comparison Group (Select one of the following)

☐ Typical intervention (typical intervention for that setting, without additions that make up the intervention being evaluated)
- ☐ Attention placebo
- ☐ Intervention element placebo
- ☐ Alternative intervention
- ☐ Pharmacotherapy
- ☐ No intervention
- ☐ Wait list/delayed intervention
- ☐ Minimal contact
- ☐ Unable to identify type of comparison

B2 Overall confidence of judgment on type of comparison group

☐ Very low (little basis)
- ☐ Low (guess)
- ☐ Moderate (weak inference)
- ☐ High (strong inference)
- ☐ Very high (explicitly stated)
- ☐ Unable to identify comparison group

B3 Counterbalancing of change agent (participants who receive intervention from a single therapist/teacher etc were counter-balanced across intervention)

☐ By change agent
☐ Statistical (analyse includes a test for intervention)
☐ Other
☐ Not reported/None NA

B4 Group equivalence established (select one of the following)

☐ Random assignment
☐ Posthoc matched set
☐ Statistical matching
☐ Post hoc test for group equivalence

B5 Equivalent mortality
☐ Low attrition (less than 20 % for post)
☐ Low attrition (less than 30% for follow-up)
☐ Intent to intervene analysis carried out?
Findings_____

Overall rating for Comparison group _2___

3= Strong Evidence 2=Promising Evidence 1=Weak Evidence 0=No Evidence

C. Appropriate Statistical Analysis

Analysis 1_____**ANOVA**_____

☐ Appropriate unit of analysis
☐ Familywise/experimenter wise error rate controlled when applicable
☐ Sufficiently large N – NO.Power analysis indicated sample should have been 94 participants

Analysis 2_____

☐ Appropriate unit of analysis
☐ Familywise/experimenter wise error rate controlled when applicable
☐ Sufficiently large N

Analysis 3_____

☐ Appropriate unit of analysis
☐ Familywise/experimenter wise error rate controlled when applicable
☐ Sufficiently large N

Overall rating for Comparison group __**2**__

3= Strong Evidence **2=Promising Evidence** **1=Weak Evidence** **0=No Evidence**

I. Follow Up Assessment

☐ Timing of follow up assessment: specify_____ 2 weeks_____

☐ Number of participants included in the follow up assessment: specify___all_____

☐ Consistency of assessment method used: specify_____TAI_____

Rating for Follow Up Assessment (select 0, 1, 2, or 3)**:** ☐ **3** ☐ **2** ☐ **1** ☐ **0**

Horner et al. (2005). The Use of Single-Subject Research to Identify Evidence-Based Practice in Special Education

Quality Indicators Within Single-Subject Research

Coding Protocol

Article Reference: Beidel, D., Turner, S., & Taylor-Ferreira, J.C. (1999). Teaching Study Skills and Test-Taking Strategies to Elementary School Students, *23*(4), 630–646.

Type of Publication: Peer reviewed Journal
Study ID: 2

Description of Participants and Setting

Participants are described with sufficient detail to allow others to select individuals with similar characteristics; (e.g., age, gender, disability, diagnosis).
Yes
No
N/A
Unknown/Unable to Code

The process for selecting participants is described with operational precision.
Yes
No
N/A
Unknown/Unable to Code

Critical features of the physical setting are described with sufficient precision to allow replication.
Yes
No
N/A
Unknown/Unable to Code

Overall Rating of Evidence: 3 2 1 0

Dependent Variable
Dependent variables are described with operational precision.
Yes
No
N/A
Unknown/Unable to Code

Each dependent variable is measured with a procedure that generates a quantifiable index.
Yes
No
N/A
Unknown/Unable to Code

Measurement of the dependent variable is valid and described with replicable

precision.
Yes
No
N/A
Unknown/Unable to Code

Dependent variables are measured repeatedly over time.
Yes
No
N/A
Unknown/Unable to Code

Data are collected on the reliability or inter-observer agreement associated with each dependent variable, and IOA levels meet minimal standards
Yes
No
N/A
Unknown/Unable to Code

Overall Rating of Evidence: 3 2 1 0

Independent Variable
Independent variable is described with replicable precision.
Yes
No
N/A
Unknown/Unable to Code

Independent variable is systematically manipulated and under the control of the experimenter.
Yes
No
N/A
Unknown/Unable to Code

Overt measurement of the fidelity of implementation for the independent variable is highly desirable.
Yes
No
N/A
Unknown/Unable to Code

Overall Rating of Evidence**:** 3 2 1 0

Baseline
The majority of single-subject research studies will include a baseline phase that provides repealed measurement of a dependent variable and establishes a pattern of responding that can be used to predict the pattern of future performance, if introduction or manipulation of the independent variable did not occur.
Yes
No
N/A
Unknown/Unable to Code

Baseline conditions are described with replicable precision.
Yes
No
N/A
Unknown/Unable to Code

Overall Rating of Evidence: 3 2 1 0

Experimental Control/internal Validity

The design provides at least three demonstrations of experimental effect at three different points in time.
Yes
No
N/A
Unknown/Unable to Code

The design controls for common threats to internal validity (e.g., permits elimination of rival hypotheses).
Yes
No
N/A
Unknown/Unable to Code

The results document a pattern that demonstrates experimental control.
Yes
No
N/A
Unknown/Unable to Code

Overall Rating of Evidence**:** 3 2 1 0

External Validity

Experimental effects are replicated across participants, settings, or materials to establish external validity.
Yes
No
N/A
Unknown/Unable to Code

Overall Rating of Evidence: 3 2 1 0

Social Validity

The dependent variable is socially important.
Yes
No
N/A
Unknown/Unable to Code

The magnitude of change in the dependent variable resulting from the intervention is socially important.
Yes
No

N/A
Unknown/Unable to Code

Implementation of the independent variable is practical and cost effective
Yes
No
N/A
Unknown/Unable to Code

Social validity is enhanced by implementation of the independent variable over extended time periods, by typical intervention agents, in typical physical and social contexts.
Yes
No
N/A
Unknown/Unable to Code

Overall Rating of Evidence: 3 2 1 0

Average WoE A across the 7 judgement areas:
Sum of X / N = 14/7 = 2
X = individual quality rating for each judgement area
N = number of judgement areas
Overall Rating of Evidence: 3 2 1 0