Case Study 1: An Evidence-Based Practice Review Report

Theme: School Based Interventions for Learning

How Effective Are Test-Taking Strategy Interventions in Improving the Academic Attainment of Secondary Aged Students with Learning Difficulties?

Summary

With the growing emphasis on assessment at secondary schools, a body of research focusing on interventions that support students in their test-taking skills, is being established. These 'test-taking strategies' have been designed to help students adopt a more organised and active approach to taking tests. Research has focused specifically on students with learning difficulties, who have a longstanding history of poor attainment and inadequate test-taking skills. This systematic literature review aims to discover how effective test-taking strategies are in improving the academic attainment of these students. Through a comprehensive search and appraisal of the literature, some promising evidence of test-taking strategy interventions emerged. However, a range of limitations indicated that findings were largely inconclusive. Implications for potential use and future directions are discussed.

Introduction

What Are Test-Taking Strategies?

Test-taking strategies (TTS) have the broad aim of making a test-taker an active participant, rather than a passive user, of school-based tests involving a sequence of questions under timed conditions (Banks & Eaton, 2014). As noted by Hughes, Maccini and Gagnon (2003), TTS can vary in their presentations and may be acquired through trial and error or incidental learning. This review however, will focus on TTS that are explicitly taught to pupils in order to increase their ability to take a systematic and organised approach when completing test-based summative assessments.

There exist a number of general TTS, such as being wary of the time, checking work carefully and answering easier questions first (Conderman & Pedersen, 2010). Research has concentrated on specific TTS taught to students in preparation for tests. Specific TTS consist of a set of steps that students follow as they complete a test and are often based on a mnemonic device, to facilitate recall of the steps of the strategy. Table 1 outlines a number of popular mnemonic-based strategies used in schools; some have been devised for mathematics or literacy related tests and others are thought to be adaptable for use in a range of school-based tests.

Modes of delivery of TTS instruction can also vary, from face-to-face small group or whole-class teaching, to delivery via multimedia platforms (Lancaster et al., 2006). An instructional sequence for teaching specific TTS was put forward by Mercer and Pullen (2005). They argued that TTS can be taught most effectively if the eight steps

shown in Table 2 are followed. This review will touch upon a range of strategies and modes of instruction, in evaluating the overall effectiveness of interventions based on TTS.

Table 1.

Examples of Test-taking Strategies Based on Mnemonic Devices

Mnemonic based strategy	Summary of steps
PIRATES A general test-taking strategy, for use in multiple choice tests (Hughes, Schumaker, Deshler & Mercer, 1988)	**Prepare to Succeed:** Pupils put their name and the mnemonic device PIRATES on the test. They make decisions about how much time to spend and in which order to complete sections of the test. Pupils make a positive statement about the testing situation, and begin the test as quickly as they can. **Inspect the Instructions:** Pupils read the instructions carefully, underline key words in the instructions, and pay attention to special requirements. **Read, Remember, and Reduce:** Pupils read each question and all of the answer choices, pause to remember what they have studied and reduce answer choices by eliminating answers that they know are incorrect. **Answer or Abandon:** Pupils answer questions that they are confident about and temporarily abandon questions that they are unsure of. **Turn Back:** Pupils go back through the test to identify each question that was previously abandoned. **Estimate:** Pupils answer the question if they know the answer or use an appropriate guessing technique only if they do not know the answer. **Survey:** Pupils check that each question is answered and that their answers are neat and legible.
SIGNS A maths test-taking strategy, for use in solving word problems (Watanabe, 1991)	**Survey question:** Pupils read the problem and underline numerals or number words. **Identify key words and labels:** Pupils look for key words that give an idea of the mathematical operation to be used, as well as labels that describe the objects being dealt with in the problem. **Graphically draw problem:** Pupils draw a picture that illustrates what the problem is asking. **Note type of operation(s) needed:** Pupils identify the operation or equation that best describes the drawing and write this down. **Solve and check problem:** Pupils compute the answer and check with a calculator If possible.
ANSWER A literacy test-taking strategy, for use in answering essay questions (Hughes, Schumaker & Deshler, 2005)	**Analyse the action words in the question:** Pupils read the question carefully and underline key words. **Notice the requirements of the questions:** Pupils mark key essay requirements and change the question into their own words. **Set up an outline:** Pupils list the main ideas of their essay within an outline format. **Work in detail:** Pupils add important details to the outline that they plan to include in their essay **Engineer your answer:** Pupils write the essay including an introductory sentence and detailed sentences about each of the main ideas in their outline. **Review your answer:** Pupils check that all parts of the question have been answered and edit their essay.

Table 2.

Steps for Teaching Mnemonic Test-taking Strategies (Mercer & Pullen, 2005)

	Step	
1	Pre-test and make commitments	Teacher provides an orientation of the strategy, administers a pre-test of the strategy, shares results others have achieved when learning the strategy, and gains pupil commitment for learning the strategy.
2	Describe the strategy	Teacher describes situations where pupils can use the strategy and presents the steps of the strategy.
3	Model the strategy	Teacher models by using the think-aloud process to demonstrate how to use the strategy. During this step, the teacher instructs pupils on what to do and how to think about each strategy step.
4	Verbal rehearsal of the strategy	Pupils describe each step of the strategy and name each step quickly in the correct order.
5	Controlled practice and feedback	Pupils practice the strategy with relatively easy examples and materials.
6	Advanced practice and feedback	Pupils practice the strategy with more difficult examples and materials.
7	Confirm acquisition and make generalisation commitments	After pupils have demonstrated mastery of the strategy, the teacher invites them to make a commitment to generalise the strategy.
8	Generalisation	Teacher facilitates discussions and provides prompts and reminders around uses of the strategy in various situations, e.g. teachers can support pupils to brainstorm how they can use the strategy in upcoming end of term exams.

Basis in Psychological Theory

The best test-takers are believed to be those who have skills beyond the content being tested; they have an understanding of the purposes and requirements of the test they are taking (Banks & Eaton, 2014). This idea of deeper thinking reflects the concept of metacognition, which involves active monitoring, regulation and orchestration of cognitive processes to achieve cognitive goals (Flavell, 1979), and which forms the foundation of many TTS interventions. At the same time, a number

of cognitive skills integral to test-taking have been identified by researchers, including general problem solving, deductive reasoning and attention to appropriate cues (Scruggs & Mastropieri, 1988). These cognitive skills have been found to be lacking in secondary aged students who have mild to moderate learning difficulties (LD); nevertheless these groups of young people may find themselves in mainstream classrooms, regularly taking tests that rely on skills with strong cognitive and metacognitive elements. Particularly in terms of test-taking situations, research has demonstrated that students with LD tend to concentrate on the wrong part of test directions, be led astray by irrelevant information and lack persistence in searching for useful information (Reid & Hresko, 1981; Scruggs & Mastropieri, 1988). Therefore, specific teaching of these cognitive and metacognitive skills might be expected to bring about positive change for these groups of young people, who may subsequently become more independent and self-regulated learners. This line of thinking led to the development of interventions based on TTS.

Relevance for Educational Psychology Practice Today

The importance of testing has been delineated in the Assessment for Learning Strategy (Department for Children, Schools and Families, 2008) as well as the SEND Code of Practice (Department for Education & Department of Health, 2014), in which teachers' assessment is made integral in the four-part cycle of support for students with LD. Furthermore, the growing usage of the Year 7 Cognitive Ability Tests across Local Authorities (Flint & Peim, 2012), adds to the picture of 'exam culture' that many secondary students will find themselves immersed in. The outcomes of such 'high-stakes testing' are used to make decisions around access to future educational

opportunities. Moreover, the link between the poor academic performance exhibited by students with LD and poor post-secondary outcomes is unfortunately, not atypical (Emerson & Hatton, 2008).

As supporting young people with LD to thrive in a range of educational settings falls within the remit of Educational Psychologists (EPs), interventions based on improving academic test performance become relevant for EP practice. Woods (2000) argued that EPs 'play a central role' in determining the GCSE assessment needs of individual candidates with LD; this could potentially involve recommendations around TTS interventions. EPs may be relevant professionals who could deliver training on specific TTS to school-based staff. Furthermore, considering the aforementioned limitations faced by students with LD, access to interventions that are found to improve attainment, could also be regarded as a matter of equal opportunities.

Review Question

In light of the above rationale, the review will evaluate the research base on TTS interventions. The review question to be answered is: *How effective are test-taking strategy interventions in improving the academic attainment of secondary aged students with learning difficulties?*

Critical Review of the Evidence Base

Literature Search

To address the current review question a comprehensive search of the most pertinent online databases (PsychINFO, ERIC and Web of Science) was carried out on 15th January 2016. Table 3 displays the search terms used in order to systematically locate the studies.

Table 3.

Search Terms Used in Databases

1. The study implemented an intervention based on test-taking strategies
2. The study was about academic attainment
3. The study involved secondary aged students
4. The study was about young people with learning difficulties

1	2	3	4
Test-taking strategy	Attainment	Secondary students	Learning di*
Test preparation	Test performance	Secondary school	Special educational
Test strategies	Quiz performance	Secondary pupils	needs
Mnemonic strategies	Academic performance	High school	SEN
		Adolescen*	Special education
		NOT college	
		NOT primary school	

**wildcard search item*

Inclusion and Exclusion Criteria

The initial search yielded 113 studies, 15 of which were removed for being duplicates. The remaining 98 studies were screened at title and abstract using the inclusion/exclusion criteria (see Table 4), leaving 8 papers to be screened at full text for further clarity. An ancestral search was then conducted on the 4 remaining papers, by firstly screening titles for appropriate terminology (for example ‘learning difficulties’, ‘adolescents’ and ‘test-taking strategies’). This search yielded 30 studies, of which a further 29 were removed for duplication and non-compliance to inclusion

criteria. In total, through systematic database and ancestral searches, 5 papers were retrieved for critical analysis. Figure 1 represents this process visually, while Appendix A details the list of studies excluded at full text screening.

A list of the final papers eligible for appraisal have been displayed in Table 5. It is notable that paper 3 (Lancaster et al., 2006) contains two studies; a pilot and experimental study. Upon review of the research procedures and results, the decision was made to exclude the findings from Lancaster et al.'s (2006) pilot study from the overall critical evaluation of the literature. This was because the methodology of the pilot study was demonstrated to be considerably flawed, consisting of only 3 participants in a pre-test post-test design, with no quantitative data on the individual participant scores. Therefore, only the experimental study in this paper was included in the critical review. The final five papers are summarised in Appendix B.

Table 4.

Inclusion and Exclusion Criteria Used for Literature Search with Rationale

Criterion	No.	Inclusion	Exclusion	Rationale
Publication type	1	Study is in a peer reviewed journal	Study is not in a peer reviewed journal	Studies in peer reviewed journals tend to be of higher quality as they have withstood some level of scrutiny by peers within the field
	2	Study contains primary empirical data	Study does not contain primary empirical data, e.g. it is a review article	Primary empirical data is first-hand information that can be collated and systematically reviewed
	3	Study is written in English	Study is not written in English	Lack of resources prevent access to translation services
	4	Study is published before 15th January 2016	Study is published on or after 15th January 2016	To include all relevant research within the topic, before a final search date that will allow time for the final review to be compiled
Participants/ Setting	5	Study involves young people at secondary school level (aged between 11 – 18)	Study does not involve young people at secondary school level, e.g. it focuses on primary aged children or young people in Higher Education	Testing becomes more substantial and consequential during the secondary school years
	6	Participants are identified as having learning difficulties	Participants are identified as typically developing or having a separate disorder, e.g. autism	This links directly to the review question; this is the population of interest
	7	Study is based in a school setting; either mainstream or special school	Study is not based in a school, e.g. it takes place in a clinic setting	The review question is concerned with school-based interventions
Intervention/ Study details	8	Study follows an experimental design, reporting quantitative data on pre- and post-measures	Study does not follow an experimental design, e.g. it has an exploratory or qualitative design	To enable effect sizes to be generated, allowing an analysis of the effectiveness of the intervention
	9	Intervention implemented is based on test-taking strategies	Intervention is not based on test-taking strategies, e.g. it focuses specifically on literacy skills	This links directly to the research question; test-taking strategy interventions are the focus of this review
	10	Outcome measures involve academic performance, e.g. maths attainment or usage of strategy during a test	Outcome measures do not involve academic performance, e.g. they focus on social/emotional factors	This links directly to the research question; the general topic of interest concerns interventions for learning

Figure 1.

Flow Diagram of the Study Selection Process

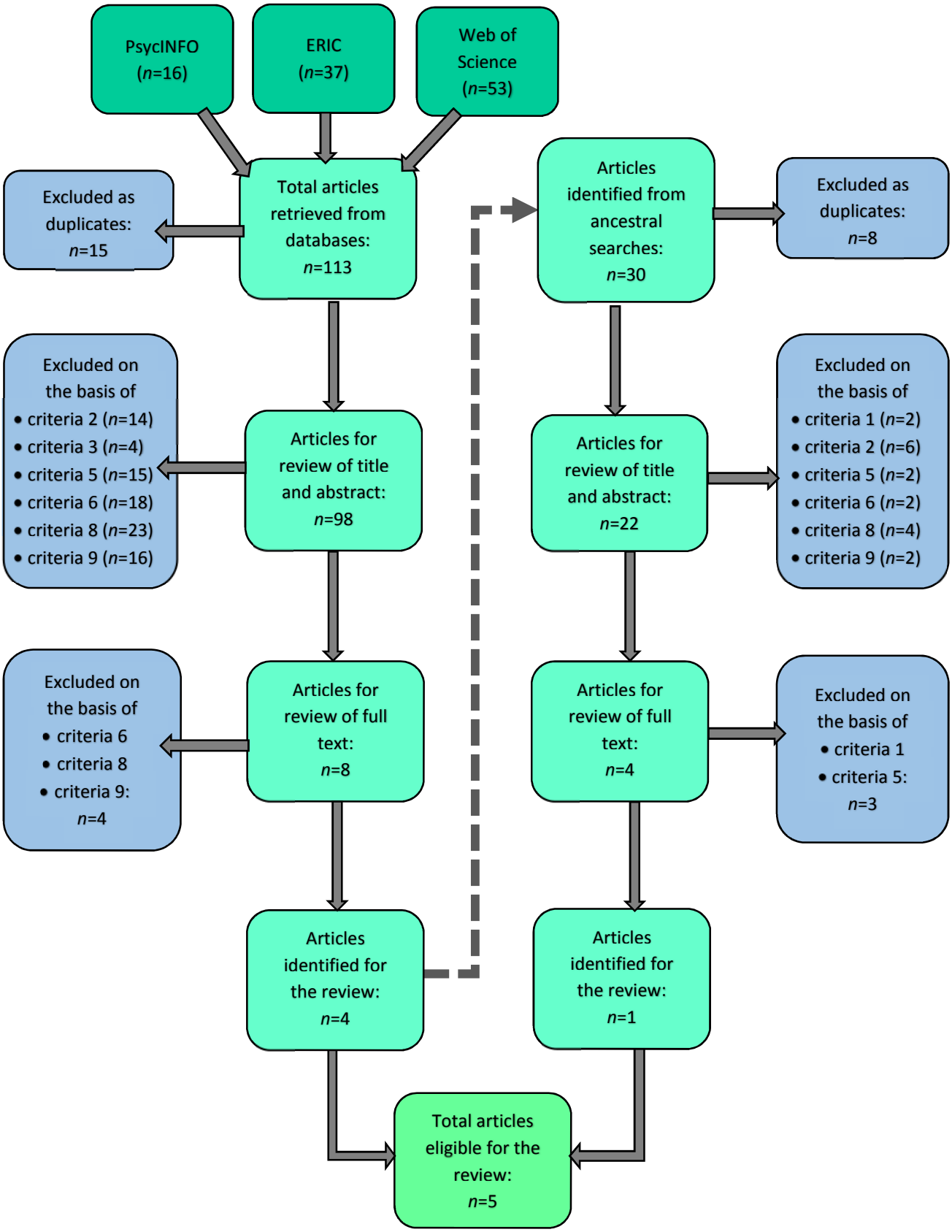

Table 5.

List of Studies Selected for Final Review

	Eligible Studies
1	Carter, E., Wehby, J., Hughes, C., Johnson, S., Plank, D., Barton-Arwood, S., & Lunsford, L. (2005). Preparing Adolescents With High-Incidence Disabilities for High-Stakes Testing With Strategy Instruction. *Preventing School Failure: Alternative Education For Children And Youth*, *49*(2), 55-62.
2	Hughes, C., & Schumaker, J. (1991). Test-taking strategy instruction for adolescents with learning disabilities. *Exceptionality*, *2*(4), 205-221.
3	Lancaster, P., Lancaster, S., Schumaker, J., & Deshler, D. (2006). The efficacy of an interactive hypermedia program for teaching a test-taking strategy to students with high-incidence disabilities. *Journal of Special Education Technology*, *21*(2), 17-30.
4	Lancaster, P., Schumaker, J., Lancaster, S., & Deshler, D. (2009). Effects of a Computerized Program on Use of the Test-Taking Strategy by Secondary Students with Disabilities. *Learning Disability Quarterly*, *32*(3), 165-179.
5	Therrien, W. J., Hughes, C., Kapelski, C., & Mokhtari, K. (2009). Effectiveness of a test-taking strategy on achievement in essay tests for students with learning disabilities. *Journal of Learning Disabilities*, *42*(1), 14-23.

Critical Comparison of Selected Studies

Weight of Evidence

The five studies selected for review were appraised using Gough's (2007) Weight of Evidence (WoE) approach. WoE is a framework used to evaluate studies based on the three areas of methodological quality (WoE A), methodological relevance (WoE B) and topic relevance (WoE C).

The WoE A for each study was determined using a coding protocol designed for evaluating methodological quality in terms of how well a study has been executed. The protocol selected for use was dependent on the research design. For single case experimental designs, the Horner et al. (2005) coding protocol was used, while

an adapted version of the APA Task Force coding protocol by Kratochwill (2003) was used for the group-based designs. Adaptations to the Kratochwill (2003) protocol are outlined in Appendix D, while full versions of completed coding protocols are in Appendices C and E.

WoE B and C are review specific, concerning the appropriateness of the methodology and topic focus of a study in relation to the review question. The weightings for each area were then averaged to provide an overall WoE for the study (WoE D). The ratings received by each study are presented below, in Table 6. The rationale for all WoE ratings can be found in Appendix F.

Table 6.

Summary of Weight of Evidence Judgements

Authors	WoE A: Quality of Methodology	WoE B: Relevance of Methodology	WoE C: Study Topic Relevance	WoE D: Overall Weight of Evidence
Carter et al. (2005)	High (2.5)	Medium (2)	Low (1)	**Medium** (1.83)
Hughes & Schumaker (1991)	High (2.5)	Medium (2)	Medium (2)	**Medium** (2.17)
Lancaster et al. (2006)	Medium (2.3)	Medium (2)	Medium (2)	**Medium** (2.1)
Lancaster et al. (2009)	Low (1.25)	Medium (2)	Medium (2)	**Medium** (1.75)
Therrien et al. (2009)	Medium (1.75)	Medium (2)	Medium (2)	**Medium** (1.91)

Participants

A total number of 208 participants were included in this review, with ages ranging from 13 to 19. The sample size in each study varied substantially, from 6 participants

(Hughes & Schumaker, 1991) to 112 participants (Lancaster et al., 2009). Crucially, all studies were underpowered. Considering that underpowered studies can 'drastically overestimate effect size estimates' (Gelman & Weakliem, 2009, p. 314), findings must be treated with caution.

All participants presented with LD, as identified by their respective US districts. It is important to note that the identification of LD is likely to differ considerably across US districts. Furthermore, there are likely to be substantial differences in the way LD is defined between the US and the UK. The fact that all participants were sampled from the USA raises significant doubts concerning the generalisability of these findings to the UK context.

Students were sampled from mainstream schools, with the majority accessing SEN support through resource bases. In the Carter et al. (2005) study, 3 participants were also identified as having a Language Impairment and 7 participants were classed as having 'Mental Retardation'. These additional needs may have impacted on the effectiveness of the intervention, however the authors failed to comment on the presence of this extraneous variable.

All of the reviewed studies provided information on the ethnic backgrounds of participants and two studies provided information on the socio-economic status of their samples. Detailed characteristics of each study's participants are displayed in Appendix B.

There appeared to have been a slight gender imbalance, with 61% of all participants being male, however this is in line with recent research that has identified 63% males make up the total population of school aged children with LD (Emerson et al., 2010). All the reviewed studies were flawed in the way they lacked information of sampling procedures. Therrien et al. (2009) stated that their participants with LD were chosen specifically for their difficulty in literacy, although their selection process was not explained.

Research Design

Two single case experimental designs (SCEDs) and three group-based designs were included in this review.

SCEDs have been argued to be 'a rigorous, scientific methodology […] proven particularly relevant for defining educational practices at the level of the individual learner' (Horner et al., 2005). The use of small sample sizes do however make generalisability difficult. The lack of control group also makes it difficult to attribute effects to the intervention.

The two reviewed SCEDs followed a multiple-probe across-participants design, which has been described as a combination of multiple baseline and probe procedures by Horner and Baer (1987). Although they were unable to be rated as 'high' on WoE B due to the aforementioned issues of SCEDs, both studies met Horner et al.'s (2005) criteria for high internal validity in the way that they

demonstrated experimental effect at least 3 times. The presence of both within-participant and inter-participant replication is reflected in the WoE ratings for both studies.

Similarly, the group-based designs were judged in relation to evidence hierarchies, which favour designs that follow a Randomised Control Trial format (Guyatt et al., 2008). Two of the three group-based studies made use of randomisation (Lancaster et al., 2009; Therrien et al., 2009), while the other used convenience methods, such as those based on timetabling. Although lacking in randomisation, Carter et al., (2005) did make use of an 'active' control group, as well as collection of data at follow-up. Because the use of an 'active' control group is more ethical than a no intervention group, which withholds a potentially beneficial intervention from a group and also because the maintenance of skills is particularly relevant for the review's population of interest, Carter et al. (2005) received a 'medium' rating on WoE B.

Additionally, group equivalence was demonstrated through statistical analyses at pre-test in all the group-based studies, which contributed to the studies' WoE A ratings.

Intervention

A range of TTS interventions were included in this review. The most common was the 'PIRATES' strategy, which was implemented in three studies. Hughes and Schumaker (1991) used a teacher-delivered approach, while Lancaster et al. (2006, 2009) used a multimedia formatted approach. As outcome measures used were different in the two approaches, it was difficult to make direct comparisons between

the two forms of delivery. Meanwhile, the 'ANSWER' strategy was implemented by Therrien et al., (2009) in their study focusing on essay tests. Carter et al. (2005) did not use a specific mnemonic device, but instead outlined an amalgamation of maths-related strategies, such as sorting and estimation. Due to the lack of specificity, this study received a 'low' rating on WoE C.

Fidelity of implementation was considered an important part in the overall evaluation of studies. Three studies overtly measured fidelity of implementation (Carter et al., 2005; Hughes & Schumaker, 1991; Therrien et al., 2009). The remaining two studies made use of a multimedia platform in delivering the TTS intervention, meaning that each participant was exposed to the same video. In spite of this seemingly equal exposure to the intervention, both Lancaster et al., (2006) and Lancaster et al., (2009) state the presence of a teacher in the room who was there to monitor and periodically check for understanding, therefore levels of support may have varied across participants. This drawback is reflected in the WoE A ratings of these studies.

Measures

The studies measured a range of academic outcomes, including attainment in literacy, maths, science and social studies. These outcomes directly relate to the review question, while others were related to learning in a more indirect way. Two studies (Lancaster et al., 2006; Lancaster et al., 2009) focused on strategy use and metacognitive outcomes, which would have been more relevant had the review question been around the ability of students with LD to *use* TTS. Additionally, Carter et al. (2005) measured the impact of a TTS intervention on academic performance

as well as test anxiety. Again, this is interesting in gaining a wider picture of test-taking for students with LD, but not directly related to the review question. These factors were considered in the studies' WoE C ratings.

The reliability of measures was commented on by only three studies (Hughes & Schumaker, 1991; Lancaster et al., 2006; Lancaster et al., 2009), who reported inter-rater agreement rates of at least 91% or Kappa value of 0.79, indicating adequate to high reliability of their particular measures (Barker, Pistrang & Elliot, 2002). Only Carter et al., (2005) detailed the process of constructing their maths attainment measure, the simulated Tennessee Competency Achievement Program (TCAP). They reported a criterion validity coefficient of 0.60 (p = 0.007), which is more than adequate, according to Barker et al. (2002). The lack of reported validity and reliability coefficients by several studies included in the review does suggest that their results should be generalised with caution.

Findings

The five studies selected for this review required different methods for calculating effect sizes. Table 7 summarises the equivalent effect sizes that should be interpreted as 'small' (questionable), 'medium' (effective) and 'large' (very effective), according to Cohen's (1988) effect size descriptors. For the majority of group-based studies, effect sizes were calculated based on the Pre-Post Control Group Standardised Mean Difference (referred to as PPC SMD). This method considers both the extent of within-group change and between-group differences, at Time 2 (T2) and Time 3 (T3), and was selected for its robustness and precision, as cited by

Morris (2007). For studies without a comparison group, Becker's (1988) Standardised Mean Difference was calculated as it crucially examines pre-test and post-test scores (referred to as PP SMD). Partial Eta Squared (η_p^2), as calculated by the study authors, has also been used as an effect size in cases where insufficient data was reported to calculate PPC SMD, such as group means and standard deviations. For SCEDs, the Percentage of Non-overlapping Data (PND) approach was used as it 'provides a good measure of treatment effectiveness' (Scruggs, Mastropieri & Castro, 1987). A summary of effect sizes for key outcomes are displayed in Table 8.

Table 7.

Indicators of a Small, Medium or Large Effect Size According to Method Used for Calculation

Type of Effect Size	Small	Medium	Large
Percentage of Non-overlapping Data points, PND (Scruggs & Mastropieri, 1998)	50-69%	70-89%	90-100%
Partial Eta Squared, η_p^2 (Cohen, 1988)	0.01	0.06	0.14
Standardised Mean Difference, SMD (Cohen, 1988)	0.20	0.50	0.80

Table 8.

Outcome	Outcome Measure	Study	WoE D	Effect Size Type	Effect Size at T2	Effect Size Descriptor	Effect Size at T3	Effect Size Descriptor

Summary of Effect Sizes

Maths attainment	TCAP	Carter et al. (2005)	Medium	PPC SMD	0.21	Small	0.37	Small-medium
Literacy attainment	Generalised Essay Measure; related to TTS	Therrien et al. (2009)	Medium	PPC SMD	0.86	Large		
	Generalised Essay Measure; not related to TTS	Therrien et al. (2009)	Medium	PPC SMD	0.16	Small		
Science and Social Studies attainment	Curriculum-based science and social studies test scores	Hughes & Schumaker (1991)	Medium	PP SMD	2.73	Large		
Strategy use	Strategy use on an essay test	Therrien et al. (2009)	Medium	PPC SMD	5.37	Large		
	Strategy use on a general knowledge test	Lancaster et al. (2009)	Medium	PPC SMD	4.76	Large		
		Hughes & Schumaker (1991)	Medium	PND	100%*	Large	100%*	Large
		Lancaster et al. (2006)	Medium	PND	100%*	Large	100%*	Large
Meta-cognition	Think Aloud	Lancaster et al. (2009)	Medium	η_p^2	0.36	Large		
		Lancaster et al. (2006)	Medium	η_p^2	0.67	Large		

**all participants in the Hughes & Schumaker (1991) and Lancaster et al. (2006) studies had a PND of 100% for the strategy use measure, meaning there was no overlap of data between Time 1 and Time 2 or Time 1 and Time 3.*

Overall, the studies were quite polarised in their demonstrated effectiveness of TTS interventions. In terms of attainment outcomes, TTS interventions were found to be very effective on measures of science and social studies (Hughes & Schumaker, 1991) as well as literacy, on essay test criterion related to the strategy that was taught (Therrien et al., 2009). These studies were both given a 'medium' rating on WoE D. At the same time, TTS interventions were found to be of questionable effect on measures of maths attainment (Carter et al., 2005) and literacy, on essay test criterion not related to the strategy taught (Therrien et al., 2009). Interestingly, these studies were also given 'medium' ratings on WoE D. This poses a quandary in answering the present review question and warrants further exploration of key outcomes.

As displayed in Table 8, the effectiveness of TTS interventions on maths attainment ranges from small at T2, to small-medium at T3. A notable change between these two points in time is that at T3, both groups had been exposed to the intervention, whereas at T2, only one group had received the intervention. It is curious that the slightly larger SMD was present after both groups had been instructed in the strategy; this difference may have been expected at T2, when only one group had been instructed in the strategy and may perhaps be indicative of a recency effect amongst participants. Carter et al. (2005) state that in comparison to regular maths instruction, there was a significant increase in maths test performance following the TTS intervention for the first group, $t(19) = 2.37$, $p < .05$, as well as for the second group, $t(17) = 2.30$, $p < .05$. They go on to argue that motivation was a substantial factor in their findings and that the effect of the intervention may have been larger if

the test given to the students was not a simulated TCAP, but the actual high-stakes TCAP assessment, which is a requirement for high school graduation.

The findings for literacy outcomes are interesting in that both small and large effect sizes were found across the two measures. Notably, Therrien et al., (2009) state that they intentionally did not provide additional TTS instruction on practice activities after students met mastery criteria on knowledge and usage of the strategy. They wanted to see if the students with LD in their sample would generalise the strategy without any prompting. They concluded that 66% of their sample were able to generalise the strategy, meaning that 34% may have required extra prompting in order to do so. This is reflected in the findings, with a larger effect size found when essays were marked in relation to the strategy taught (such as according to idea, content and organisation) and a smaller effect size found when essays were marked against criteria unrelated to the strategy.

As displayed in Table 8, a particularly large effect size was found for the science and social studies outcomes. Hughes and Schumaker (1991) report that students went from an average grade of F (a failing grade) before the TTS intervention, to an average grade of C, after being instructed on the TTS. This appears to be an impressive result and suggests generalisation of the strategy. However, there was no presence of a control group meaning that Hughes and Schumaker (1991) did not adequately control for threats to internal validity, therefore it is very difficult to attribute this improvement to the TTS intervention alone. Furthermore, despite their high inter-rater reliability for the strategy use measure, Hughes and Schumaker

(1991) failed to provide a reliability coefficient for the mainstream science and social studies tests, which would have been constructed and marked by a range of different teachers. Fundamentally, the findings of this 6-participant study illustrate Gelman and Weakliem's (2009) assertion that underpowered studies can result in exaggerated effects.

Although not directly linked to attainment, researchers such as Therrien et al. (2009) have argued that a key component of looking at the effectiveness of TTS interventions is to ascertain whether the strategy was learned and applied, which helps to determine whether gains in their performance can be attributed to the strategy. Several studies included in the review measured strategy use and metacognition. As displayed in Table 8, consistently large effects were found across these measures. The homogeneity of effect sizes calculated across the SCEDs does however pose an issue for the present review. Both Hughes and Schumaker (1991) and Lancaster et al., (2006) used the 'percentage of strategic responses performed correctly' as a measure in their multiple-probe designs. 100% PND was calculated for all participants in both studies, which appears to demonstrate high effectiveness of the intervention. However, it can be argued that focusing on how much the strategy was used over time, with increasing exposure to the strategy, makes the lack of overlapping points self-evident. These studies essentially indicate that students with LD *can learn* a TTS, but the representation taken alone does not quite capture the nature of the effect of the intervention on the review's population of interest.

Conclusion and Recommendations

This review examined two SCEDs and three group-based studies, in order to explore the effectiveness of TTS interventions on the attainment of students with LD. The reviewed studies appeared to provide promising evidence to support the effectiveness of TTS interventions, with large effects found on measures of literacy, science and social studies. Small and small-medium effects were also found on measures of maths attainment. The findings were however caveated, due to the fact that all studies received only a 'medium' overall WoE rating.

The methodological issues discussed in previous sections greatly limit the evidence supporting TTS interventions. The distinct lack of consistently gathered attainment data makes it difficult to draw any solid conclusions, as does the absence of recent research originating from the UK. Furthermore, the same research group were involved in a number of the studies, including Hughes, Lancaster and Schumaker, some of whom devised the original 'PIRATES' test-taking strategy. This undermines the evidence as these researchers may have had a vested interest in promoting the effectiveness of the intervention, leading to potential implications of publication bias.

Nevertheless, the finding that students with LD are able to learn and apply TTS that subsequently helped them to improve their writing, maths and science under test conditions without the extra assistance of adults, suggests that these interventions are worthy of further exploration by EPs. It is clear that focusing on TTS alone is unlikely to produce meaningful improvements in attainment, therefore one must look to the interactive factors at play in test-taking situations. The literature identified key

affective factors, including test anxiety and motivation; both of which should be considered in the design and implementation of TTS interventions. Rather than TTS being taught in the eleventh-hour, Banks and Eaton (2014) suggest that earlier introduction, perhaps at primary level, will be beneficial and lead to enhanced 'ownership and automaticity of skills' (p. 209). Furthermore, the revised national curriculum suggests that assessment is become more probing, therefore students with LD may benefit from not just following certain strategies but also learning to communicate what they know in a variety of ways.

Considering the paucity of rigorous research in this area, there is scope for more extensive work to add to the evidence base. Recommendations for further research include the collection of follow-up data, randomised control trials and larger sample sizes for enhanced power. Crucially, future studies are recommended to use a wider range of attainment related outcomes, preferably those that students with LD in the UK will encounter in their secondary school years.

References

Banks, T., & Eaton, I. (2014). Improving Test-Taking Performance of Secondary At-Risk Youth and Students With Disabilities. *Preventing School Failure: Alternative Education for Children and Youth*, *58*(4), 207-213.

Barker, C., Pistrang, N., & Elliott, R. (2002). *Research methods in clinical psychology: An introduction for students and practitioners*. (2nd ed) Chichester: John Wiley & Sons LTD.

Becker, B. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology*, *41*(2), 257-278.

Brannen, J. (1992). *Mixing methods*: *Qualitative and quantitative research*. Aldershot: Avebury.

Carter, E., Wehby, J., Hughes, C., Johnson, S., Plank, D., Barton-Arwood, S., & Lunsford, L. (2005). Preparing Adolescents With High-Incidence Disabilities for High-Stakes Testing With Strategy Instruction. *Preventing School Failure: Alternative Education For Children And Youth*, *49*(2), 55-62.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. New York: Lawrence Erlbaum Associates.

Conderman, G., & Pedersen, T. (2010). Preparing students with mild disabilities for taking state and district tests. *Intervention in School and Clinic*, *45*(4), 232-241. Department for Children, Schools and Families (DCSF). (2008). *The Assessment for Learning Strategy*. Nottingham: DCSF.

Department for Education and Department of Health (DfE & DH). (2014). *Special educational needs and disability code of practice: 0 to 25 years.* London: The Stationery Office.

Emerson, E., & Hatton, C. (2008). *People with Learning Disabilities in England 2008.* Lancaster: Centre for Disability Research

Emerson, E., Hatton, C., Robertson, J., Roberts, H., Baines, S., & Glover, G., (2010) *People with Learning Disabilities in England 2010.* Durham: Improving Health and Lives: Learning Disabilities Observatory

Flavell, J. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, *34*(10), 906.

Flint, K., & Peim, N. (2012). *Rethinking the education improvement agenda: A critical philosophical approach*. London: Continuum.

Gelman, A., & Weakliem, D. (2009). Of beauty, sex and power: Too little attention has been paid to the statistical challenges in estimating small effects. *American Scientist*, *97*(4), 310-316.

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

Guyatt, G., Oxman, A., Vist, G., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schünemann, H. (2008). GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ (Clinical research ed.)*.

Horner, R., & Baer, D. (1978). Multiple-probe technique: A variation of the multiple baseline. *Journal of Applied Behavior Analysis, 11,* 189-196.

Horner, R., Carr, E., Halle, J., Mcgee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, *71*(2), 165-179.

Hughes, C., Maccini, P., & Gagnon, J. (2003). Interventions that positively impact the performance of students with learning disabilities in secondary general education classes. *Learning Disabilities*, *12*(3), 101-111.

Hughes, C., & Schumaker, J. (1991). Test-taking strategy instruction for adolescents with learning disabilities. *Exceptionality*, *2*(4), 205-221.

Hughes, C., Schumaker, J., Deshler, D., & Mercer, C. (1988). *The test taking strategy.* Lawrence, KS: Excellent Enterprises.

Hughes, C., Schumaker, J., & Deshler, D. (2005). *The essay test-taking strategy: Instructor's manual*. Lawrence, KS: Edge Enterprises, Inc.

Kratochwill, T. R. (2003). Evidence-Based Practice: Promoting Evidence-Based Interventions in School Psychology. *School Psychology Quarterly, 18*(4), 389-408.

Lancaster, P., Lancaster, S., Schumaker, J., & Deshler, D. (2006). The efficacy of an interactive hypermedia program for teaching a test-taking strategy to students with high-incidence disabilities. *Journal of Special Education Technology*, *21*(2), 17-30.

Lancaster, P., Schumaker, J., Lancaster, S., & Deshler, D. (2009). Effects of a Computerized Program on Use of the Test-Taking Strategy by Secondary Students with Disabilities. *Learning Disability Quarterly*, *32*(3), 165-179.

Mercer, C., & Pullen, P. (2005). *Students with learning disabilities* (6th ed.). Upper Saddle River, NJ: Merrill.

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

Reid, D., & Hresko, W. (1981). *A cognitive approach to learning disabilities.* New York: McGraw-Hill.

Scruggs, T., & Mastropieri, M. (1988). Are learning disabled students "test-wise"?: A review of the recent research. *Learning Disabilities Focus, 3*, 87–97.

Scruggs, T., & Mastropieri, M. (1998). Summarizing single-subject research issues and applications. *Behavior Modification*, *22*(3), 221-242.

Scruggs, T., Mastropieri, M., & Casto, G. (1987). The quantitative synthesis of single-subject research methodology and validation. *Remedial and Special education*, *8*(2), 24-33.

Therrien, W. J., Hughes, C., Kapelski, C., & Mokhtari, K. (2009). Effectiveness of a test-taking strategy on achievement in essay tests for students with learning disabilities. *Journal of Learning Disabilities*, *42*(1), 14-23.

Watanabe, A. (1991). *The effects of a mathematical word problem solving strategy on problem solving performance by middle school students with mild disabilites*. Unpublished doctorial dissertation, University of Florida, Gainsville

Woods, K. (2000). Assessment Needs in GCSE Examinations: some student perspectives. *Educational Psychology in Practice*, *16*(2), 131-140.

Appendices

Appendix A

Articles excluded at full text screening from database searches

Excluded studies	Rationale for exclusion
Harris, M. L., Schumaker, J. B., & Deshler, D. D. (2011). The Effects Of Strategic Morphological Analysis Instruction On The Vocabulary Performance Of Secondary Students With And Without Disabilities. *Learning Disability Quarterly, 34*(1), 17	Criteria 9: The intervention used in this study was not based on test-taking strategies.
Hong, E., Sas, M., & Sas, J. C. (2006). Test-taking strategies of high and low mathematics achievers. *The Journal of Educational Research*, *99*(3), 144-155.	Criteria 8: This study did not follow an experimental design and instead generated qualitative data through interviews.
Ritter, S., & Idol-Maestas, L. (1986). Teaching middle school students to use a test-taking strategy. *The Journal of Educational Research*, *79*(6), 350-357.	Criteria 6: This study included both students with and without learning difficulties, however the results did not differentiate between the two populations.
Xu, Y., & Wu, Z. (2012). Test-taking strategies for a high-stakes writing test: An exploratory study of 12 Chinese EFL learners. *Assessing Writing*, *17*(3), 174-190.	Criteria 6 and criteria 8: The bilingual participants in this study were not identified as having learning difficulties. This was also an exploratory study that did not generate quantitative data.

Articles excluded at full text screening from ancestral searches

Excluded studies	Rationale for exclusion
Lee, P., & Alley, G. R (1981). *Training junior high school LD students to use a test-taking strategy* (Report No. URLD-RR-38). Lawrence, KS: Institute for Research in Learning Disabilities.	Criteria 1: This article was not published in a peer-reviewed journal.
Scruggs, T. E., & Mastropieri, M. A. (1986). Improving the test-taking skills of behaviourally disordered and learning disabled children. *Exceptional Children, 53*(1), 63-68.	Criteria 5: The participants in this study were primary aged rather than secondary aged.
Scruggs, T. E., Mastropieri, M. A., & Tolfa-Veit, D. (1986). The effects of coaching on the standardized test performance of learning disabled and behaviorally disordered students. *Remedial and Special Education*, *7*(5), 37-41.	Criteria 5: The participants in this study were primary aged rather than secondary aged.

Appendix B

Mapping the Field – Summary of Final Studies

Authors	Study Design	Sample characteristics	Intervention details	Measures	Key findings
Carter, Wehby, Hughes, Johnson, Plank, Barton-Arwood & Lunsford (2005) Study ID: 1	Group-based design: two groups, one of which was a delayed treatment group.	• Total sample size: 38 • Gender: Male (*n*=22), Female (*n*=16) • Age: range 15-19 years • Presenting difficulty: Learning disability (*n*=28), Mental retardation (*n*=7), Language impairment (*n*=3) • Ethnicities: African American (*n*=26), White (*n*=11), Hispanic (*n*=1) • Setting: All in one public high school in a 'large urban district' in Tennessee, USA	Non-specific TTS intervention: Various strategies for multiple choice maths and language arts tests/exams, e.g. sorting, eliminating, underlining. Participants (in groups of 5 to 7) received 6 sessions of the intervention, each session lasting 90 minutes	• Simulated Tennessee Competency Achievement Program-Mathematics (TCAP) - Maths multiple choice test. Administered at Time 1, Time 2 and Time 3 to all participants • Test Anxiety Inventory (TAI) – a self-report scale for test-related anxiety. Administered at Time 1 and Time 3, to all participants	The group who had received the TTS intervention first, showed significant improvement in their TCAP scores. The second group, following the delayed intervention, also showed a significant improvement in their TCAP scores. The second group also demonstrated a small but significant decrease in test anxiety, following the TTS intervention. The authors concluded that an exclusive focus on TTS alone is unlikely to produce socially valid improvements in the students' test performance and that preparation for high-stakes testing should be addressed much earlier in the students' academic programs.
Hughes & Schumaker (1991) Study ID: 2	Single case study: Multiple-probe design, with all participants experiencing three conditions	• Total sample size: 6 • Gender: Male (*n*=5), Female (*n*=1) • Age: range 13.1-17.2 years (M=15.1) • Presenting difficulty: All had a Learning Disability (formally classified according to state guidelines) • Ethnicities: African American (*n*=3), White (*n*=3) • Setting: All in mainstream schools, attending a resource base for 1 to 2 periods per day. All in Florida, USA	Specific TTS intervention: PIRATES mnemonic device (see Table 1). All participants (in groups of 3 to 5) received 18 instructional sessions, each session lasting 20 minutes	• Probe tests: general knowledge tests with true/false items and multiple choice questions. Administered at Time 1, Time 2 and Time 3, to all participants • Test-taking checklists, to assess whether participants were following the strategy steps when taking the probe tests. Administered at Time 1, Time 2 and Time 3, to all participants • Mainstream science and social studies test scores and test papers, to measure generalisation of strategies. Collected at Time 1 and Time 2	The sample were capable of successfully mastering a relatively complex TTS that comprised a sequenced set of steps, according to the multiple-probe design. Students also showed that they used the strategy up to 11 weeks after the intervention period had ended. Students also showed improvement in their mainstream science and social studies test scores, which suggests evidence of the students generalising the strategy to their regular classes.

Mapping the Field – Summary of Final Studies (continued)

Authors	Study Design	Sample characteristics	Intervention details	Measures	Key findings
Lancaster, Lancaster, Schumaker & Deshler (2006) Study ID: 3	Single case study: Multiple-probe design, with all participants experiencing three conditions	• Total sample size: 12 • Gender: Male (*n*=6), Female (*n*=6) • Age: M = 15 years 2 months • Presenting difficulty: All with a Learning Disability (as identified by their state district) • Ethnicities: African American (*n*=5), White (*n*=6), Hispanic (*n*=1) • Setting: All in mainstream high school, receiving at least 1 hour of special education services, in a Midwestern city, USA	Specific TTS intervention in a CD format: Parts of the Test-Taking Strategy Manual (Hughes et al., 1988) were converted into a multimedia format, with video recordings of student actors instructing on the PIRATES strategy. Each participant completed the CD in 3 to 5 sessions. Each session lasted 30 to 45 minutes	• Strategy use tests: general knowledge tests with multiple choice items. Administered to all participants at Time 1, Time 2 and Time 3 • Think-Aloud test was used to measure how well participants used the strategy and was administered at Time 1 and Time 2 • A strategy knowledge test was administered to assess participants' understanding and retention of the strategy. Administered at Time 1 and Time 2 • A student satisfaction questionnaire was administered post intervention to all participants, to measure social validity • A structured student interview was conducted post intervention to gather student perspectives on the intervention	All participants' use of the strategy increased substantially during and after instruction compared to the baseline probes. Specifically, all 12 students in the experimental study reached a mastery level of 90% of strategy steps used on their second practice attempt, following exposure to the TTS CD. Most of the students were able to demonstrate, through the Think Aloud Test, that they understood the steps of the strategy and how and why to use them. Student satisfaction data also indicated that students were satisfied with the CD format, but that revisions were necessary regarding the graphics and animation.
Lancaster, Schumaker, Lancaster & Deshler (2009) Study ID: 4	Group-based design with two experimental groups and two control groups (random allocation to conditions)	• Total sample size: 112 • Gender: Male (*n*=67), Female (*n*=45) • Age: specific ages not provided, but all participants of junior and senior high school age • Presenting difficulty: All with a Learning Disability (as identified by their state district) • Ethnicities: African American (*n*=38), Asian American (*n*=2), White (*n*=60), Hispanic (*n*=12) • Setting: Junior and senior high schools across two school districts in small Midwestern towns, USA	Specific TTS intervention in a CD format: Parts of the Test-Taking Strategy Manual (Hughes et al., 1988) were converted into a multimedia format, with video recordings of student actors instructing on the PIRATES strategy. The experimental groups worked for 4 sessions of 30-45 minutes each using the Test-Taking Strategy CD. The control groups worked through a different computerised program targeting self-advocacy skills for the same amount of time.	• Strategy use tests: general knowledge tests with multiple choice items. Administered to all participants at Time 1 and Time 2 • Think-Aloud test was used to measure how well participants used the strategy and was administered at Time 1 and Time 2 • A strategy knowledge test was administered to assess participants' understanding and retention of the strategy. Administered at Time 1 and Time 2 • A student satisfaction questionnaire was administered post intervention to all participants, to measure social validity • A structured student interview was conducted post intervention to gather student perspectives on the intervention • A questionnaire was also administered to a community panel, to gain their perspectives	The pre-test post-test comparison-group design demonstrated that the students who had received the intervention were significantly more able to use and explain the strategy, in comparison to the students who received a different computerised program. Student satisfaction data also indicated that students were satisfied with the CD format, but that revisions were necessary regarding the graphics and animation. The authors concluded that the TTS CD is an effective tool for teaching junior- and senior-high school students with LD a complex test-taking strategy in a large-group setting.

Mapping the Field – Summary of Final Studies (continued)

Authors	Study Design	Sample characteristics	Intervention details	Measures	Key findings
Therrien, Hughes, Kapelski & Mokhtari (2009) Study ID: 5	Group-based design, with one experimental group, one control group and a separate comparison group of typically developing students at post-test (stratified random assignment to groups)	• Total sample size: 40 (50 including the typically developing comparison group). • Gender: Male (*n*=26), Female (*n*=14) • Age: M=13.04 • Presenting difficulty: 40 identified as Learning Disabled, with a specific difficulty in reading and/or written expression • Ethnicities: White American (*n*=37), Hispanic (*n*=3) • Setting: Mainstream schools in rural Ohio, USA	Specific TTS intervention: Essay test-taking strategy - ANSWER (see Table 1) Experimental group (in groups of 7) received 8 sessions of TTS instruction, lasting 42 minutes each, for over 2 weeks	• Persuasive essay questions were given and evaluated at Time 1 and Time 2. These were marked according to a generalised essay measure to assess 6 traits of essay writing. • Some of the essay measure criterion were related to the TTS – these were used to evaluate the essays based on the TTS, while others were unrelated to the TTS – these were used to see if the TTS intervention had generalised to other aspects of writing	The intervention group scored significantly higher than the control group on ratings related to the TTS, e.g. organisation and idea quality. There was no significant difference between the two groups on the essay measure that didn't relate to the TTS. Students in the typically developing comparison group scored higher on the essay measure compared to the students with LD. The authors conclude that the ANSWER strategy can be effective in improving the essay writing performance of students with LD, however many students may require additional instruction before they are able to apply the strategy beyond the instructional setting.

Appendix C

Coding Protocols

Single case design:
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. *Exceptional Children*, *71*(2), 165-179.

Study ID Number: 3

Research Design: Single case experimental design (multiple-probe)

Name of Coder: **Date:** 21.01.16

Full Study Reference: Lancaster, P. E., Lancaster, S. J., Schumaker, J. B., & Deshler, D. D. (2006). The efficacy of an interactive hypermedia program for teaching a test-taking strategy to students with high-incidence disabilities. *Journal of Special Education Technology*, *21*(2), 17-30.

Intervention Name (description of study): Computerised test-taking strategy instruction

---

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

1. **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 *Also included SES related information*
☐No
☐N/A
☐Unknown/unable to code

The process for selecting participants is described with replicable precision
☐Yes
☒No *The process of recruitment is not described*
☐N/A
☐Unknown/unable to code

Critical features of the physical setting are described with sufficient precision to allow replication
☒Yes *Information is provided on the general school district as well as the specific instructional setting*
☐No
☐N/A
☐Unknown/unable to code

Overall Rating of Evidence: ☐3 ☒2 ☐1 ☐0

This score was given because participant and setting details are provided, but no information is given on the process of participant recruitment.

2. Dependent Variable

Dependent variables are described with operational precision
☒Yes *Probe tests, questionnaires and interviews are described in some detail*
☐No
☐N/A
☐Unknown/unable to code

Each dependant variable is measured with a procedure that generates a quantifiable index.
☐Yes
☒No *Interviews generated qualitative data*
☐N/A
☐Unknown/unable to code

Measurement of the dependant variable is valid and described with replicable precision.
☐Yes
☒No *Insufficient detail is provided for some measures, e.g. Think Aloud*
☐N/A
☐Unknown/unable to code

Dependant variables are measured repeatedly over time
☒Yes *Strategy Use is measured repeatedly*
☐No
☐N/A
☐Unknown/unable to code

Data are collected on the reliability or inter-observer agreement associated with each dependant variable, and IOA levels meet minimal standards
☒Yes *Inter-scorer reliability is 91% - 98% across the Strategy Use tests, Strategy Knowledge test and Think Aloud tests*
☐No
☐N/A
☐Unknown/unable to code

Overall Rating of Evidence: ☐3 ☒2 ☐1 ☐0
This score was given because there is a lack of detail around some measures, e.g. the metacognitive Think Aloud test

3. Independent Variable

Independent variable is described with replicable precision
☒Yes *The computerised test-taking instruction is described in some detail*
☐No
☐N/A
☐Unknown/unable to code

Independent variable is systematically manipulated and under the control of the experimenter

☒Yes *Intervention was delivered systematically*
☐No
☐N/A
☐Unknown/unable to code

Overt measurement of the fidelity of implementation for the independent variable is highly desirable
☐Yes
☒No *the same CD was shown to each participant however overt measurement of fidelity was not mentioned*
☐N/A
☐Unknown/unable to code

Overall Rating of Evidence: ☐3 ☒2 ☐1 ☐0
This score was given because the Independent variable was manipulated in a structured way however there was no overt measurement of the fidelity of implementation

4. Baseline

The design provides a baseline phase that gives repeated measurements (3+) of a dependent variable
☒Yes *There are 3+ baseline measures of Strategy Use*
☐No
☐N/A
☐Unknown/unable to code

Baseline conditions are described with replicable precision
☒Yes *Described on page 23*
☐No
☐N/A
☐Unknown/unable to code

Overall Rating of Evidence: ☒3 ☐2 ☐1 ☐0
This score was given because baseline conditions are described and 3+ measures were given to each participant

5. Experimental Control/Internal Validity

The design provides at least three demonstrations of experimental effect at three different points in time
☒Yes *Experimental effects are demonstrated at baseline, practice and post intervention*
☐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 *Variance in teaching was not controlled for*
☐N/A

☐Unknown/unable to code

The results document a pattern that demonstrates experimental control
☒Yes *There was no overlap between baseline and intervention periods*
☐No
☐N/A
☐Unknown/unable to code

Overall Rating of Evidence: ☐3 ☒2 ☐1 ☐0
This score was given because experimental effect was demonstrated, however extraneous variables may be impacting on these findings

6. Social validity

Experimental effects are replicated across participants, settings, or materials to establish external validity
☒Yes *Experimental effects were found across all participants*
☐No
☐N/A
☐Unknown/unable to code

Selection and attribution biases (e.g., the selection of only certain participants, or the publication of only successful examples) are minimized
☐Yes
☐No
☐N/A
☒Unknown/unable to code

The dependent variable is socially important
☒Yes *Acquisition of a computerised test-taking strategy has implications for large-scale intervention development*
☐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 *Costing information of the CD is not provided*

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

This score was given because the effect was found for all participants irrespective of gender, age, presenting difficulty and ethnicity, which is the key criteria for a high weighting as identified by Horner (2005)

Average WoE A across the 6 judgement areas:

Sum of X / N = 14/6 = 2.3

X = individual quality rating for each judgement area
N = number of judgement areas

Overall Rating of Evidence: ☐3 ☒2 ☐1 ☐0

Appendix D

Amendments to the Kratochwill (2003) Coding Protocol, with rationale

Amendment	Rationale
Section I, B7 and B8 – qualitative research methods removed	All studies generated quantitative data, as per the inclusion criteria
Section II, C - primary/secondary outcomes removed	Outcomes will be considered separately (looking at effect sizes and tabulating with WoE D) as part of the review process
Section II, D – educational/clinical significance removed	This section relates to outcomes which are dealt with separately in this review
Section II, E – identifiable components removed	This section relates to primary/secondary outcomes, which have previously been excluded
Section II, G – replication removed	Not relevant for current review question
Section II, H - site of implementation (including H1 and H2) removed	All studies were conducted in schools, as per inclusion criteria, so site of implementation does not need to be considered
Section III, A2 – participant characteristics removed	This information has already been gathered and displayed in the 'map the field' table
Section III, D – dosage removed	Not relevant for current review question
Section III, I – training/support removed	Staff implementing the interventions in the studies were not typically school staff

Appendix E

Coding Protocols

Group-based design:

Kratochwill, T. R. (2003). Evidence-Based Practice: Promoting Evidence-Based Interventions in School Psychology. *School Psychology Quarterly, 18*(4), 389-408.

Domain:

☐School- and community-based intervention programs for social and behavioral problems
☒Academic intervention programs
☐Family and parent intervention programs
☐School-wide and classroom-based programs
☐Comprehensive and coordinated school health services

Name of Coder(s): **Date**: 01/23/16
M / D / Y

Full Study Reference in APA format: Lancaster, P., Schumaker, J., Lancaster, S., & Deshler, D. (2009). Effects of a Computerized Program on Use of the Test-Taking Strategy by Secondary Students with Disabilities. *Learning Disability Quarterly*, *32*(3), 165-179.

Intervention Name (description from study): Computerised test-taking strategy instruction

Study ID Number (Unique Identifier): 4

Type of Publication: (Check one)

☐Book/Monograph
☒Journal article
☐Book chapter
☐Other (specify):

I. General Characteristics

A. General Design Characteristics

A1. Random assignment designs (if random assignment design, select one of the following)
- A1.1 ☒Completely randomized design
- A1.2 ☐Randomized block design (between-subjects variation)
- A1.3 ☐Randomized block design (within-subjects variation)
- A1.4 ☐Randomized hierarchical design

A2. Nonrandomized designs (if nonrandom assignment design, select one of the following)
- A2.1 ☐Nonrandomized design
- A2.2 ☐Nonrandomized block design (between-participants variation)
- A2.3 ☐Nonrandomized block design (within-participants variation)
- A2.4 ☐Nonrandomized hierarchical design
- A2.5 ☐Optional coding of Quasi-experimental designs (see Appendix C)

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

A3.1 ☐Very low (little basis)
A3.2 ☐Low (guess)
A3.3 ☐Moderate (weak inference)
A3.4 ☒High (strong inference)
A3.5 ☐Very high (explicitly stated)
A3.6 ☐N/A
A3.7 ☐Unknown/unable to code

B. Statistical Treatment/Data Analysis (answer B1 through B6)

B1. Appropriate unit of analysis ☒yes ☐no

B2. Familywise error rate controlled ☐yes ☒no ☐N/A

B3. Sufficiently large *N* ☐yes ☒no
Statistical Test: repeated measures ANOVA
_ level: 0.05
ES: large
N required: 64 in each group

B4. Total size of sample (start of the study): 112
N

B5. Intervention group sample size: 58
N

B6. Control group sample size: 54
N

~~**For studies using qualitative research methods, code B7 and B8**~~

~~B7. Coding~~

~~B7.1 Coding scheme linked to study's theoretical-empirical basis (select one) ☐yes ☐no~~

~~B7.2 Procedures for ensuring consistency of coding are used (select one) ☐yes ☐no~~

~~Describe procedures: _____~~

~~B7.3 Progression from abstract concepts to empirical exemplars is clearly articulated (select one) ☐yes ☐no~~

~~B8. Interactive process followed (select one) ☐yes ☐no~~

~~Describe process: _____~~

C. Type of Program (select one)

C1. ☐Universal prevention program
C2. ☐Selective prevention program
C3. ☐Targeted prevention program
C4. ☒Intervention/Treatment
C5. ☐Unknown

D. Stage of the Program (select one)
D1. ☐Model/demonstration programs
D2. ☒Early stage programs
D3. ☐Established/institutionalized programs
D4. ☐Unknown

E. Concurrent or Historical Intervention Exposure (select one)
E1. ☐Current exposure
E2. ☐Prior exposure
E3. ☒Unknown

II. Key Features for Coding Studies and Rating Level of Evidence/ Support
(3=Strong Evidence 2=Promising Evidence 1=Weak Evidence 0=No Evidence)

A. Measurement (answer A1 through A4)

A1. Use of outcome measures that produce reliable scores for the majority of primary outcomes. The table for Primary/Secondary Outcomes Statistically Significant allows for listing separate outcomes and will facilitate decision making regarding measurement (select one of the following)
A1.1 ☒Yes
A1.2 ☐No
A1.3 ☐Unknown/unable to code

A2. Multi-method (select one of the following)
A2.1 ☒Yes
A2.2 ☐No
A2.3 ☐N/A
A2.4 ☐Unknown/unable to code

A3. Multi-source (select one of the following)
A3.1 ☒Yes
A3.2 ☐No
A3.3 ☐N/A
A3.4 ☐Unknown/unable to code

A4. Validity of measures reported (select one of the following)
A5.1 ☐Yes validated with specific target group
A5.2 ☐In part, validated for general population only
A5.3 ☒No
A5.4 ☐Unknown/unable to code

Rating for Measurement (select 0, 1, 2, or 3)**:** ☐**3** ☐**2** ☒**1** ☐**0**

B. Comparison Group

B1. Type of Comparison Group (select one of the following)
B1.1 ☐Typical contact
B1.2 ☐Typical contact (other) specify:

B1.3 ☐Attention placebo
B1.4 ☒Intervention elements placebo
B1.5 ☐Alternative intervention
B1.6 ☐PharmacotherapyB1.1
B1.7 ☐No intervention
B1.8 ☐Wait list/delayed intervention
B1.9 ☐Minimal contact
B1.10 ☐Unable to identify comparison group

Rating for Comparison Group (select 0, 1, 2, or 3): ☐3 ☒2 ☐1 ☐0

B2. Overall confidence rating in judgment of type of comparison group (select one of the following)
B2.1 ☐Very low (little basis)
B2.2 ☐Low (guess)
B2.3 ☒Moderate (weak inference)
B2.4 ☐High (strong inference)
B2.5 ☐Very high (explicitly stated)
B2.6 ☐Unknown/Unable to code

B3. Counterbalancing of Change Agents (answer B3.1 to B3.3)
B3.1 ☐By change agent
B3.2 ☐Statistical
B3.3. ☒Other

B4. Group Equivalence Established (select one of the following)
B4.1 ☐Random assignment
B4.2 ☐Posthoc matched set
B4.3 ☐Statistical matching
B4.4 ☒Post hoc test for group equivalence

B5. Equivalent Mortality (answer B5.1 through B5.3)
B5.1 ☒Low Attrition (less than 20% for Post)
B5.2 ☐Low Attrition (less than 30% for follow-up)
B5.3 ☐Intent to intervene analysis carried out

Findings:_____

~~**C. Primary/Secondary Outcomes Are Statistically Significant**~~

~~C1. Evidence of appropriate statistical analysis for **primary outcomes** (answer C1.1 through C1.3)~~
~~C1.1 ☐Appropriate unit of analysis (rate from previous code)~~
~~C1.2 ☐Familywise/experimenterwise error rate controlled when applicable (rate from previous code)~~
~~C1.3 ☐Sufficiently large *N* (rate from previous code)~~

~~C2. Percentage of **primary outcomes** that are significant (select one of the following)~~
~~C2.1 ☐Significant primary outcomes for at least 75% of the total primary outcome measures for each key construct~~
~~C2.2 ☐Significant primary outcomes for between 50% and 74% of the total primary outcome measures for each key construct~~

C2.3 ☐Significant primary outcomes for between 25% and 49% of the total primary outcome measures for any key construct

Rating for Primary Outcomes Statistically Significant (select 0, 1, 2, or 3)**: ☐3 ☐2 ☐1 ☐0**

C3. Evidence of appropriate statistical analysis for **secondary outcomes** (answer C3.1 through C3.3)

C3.1 ☐Appropriate unit of analysis

C3.2 ☐Familywise/experimenterwise error rate controlled when applicable (rate from previous code)

C3.3 ☐Sufficiently large *N* (rate from previous code)

C4. Percentage of **secondary outcomes** that are significant (select one of the following)

C4.1 ☐Significant secondary outcomes for at least 75% of the total secondary outcome measures for each key construct

C4.2 ☐Significant secondary outcomes for between 50% and 74% of the total secondary outcome measures for each key construct

C4.3 ☐Significant secondary outcomes for between 25% and 49% of the total secondary outcome measures for any key construct

Rating for Secondary Outcomes Statistically Significant (select 0, 1, 2, or 3)**: ☐3 ☐2 ☐1 ☐0**

C5. Overall Summary of Questions Investigated

C5.1 ☐Main effect analyses conducted (select one) yes no

C5.2 ☐Moderator effect analyses conducted (select one) yes no

Specify results: _____

C5.3. ☐Mediator analyses conducted (select one) yes no

Specify results: _____

C. Primary/Secondary Outcomes Statistically Significant (only list $p \leq .05$)

(list primary outcomes first in alphabetical order, followed by secondary outcomes in alphabetical order)

Outcomes	Primary vs. Secondary	Who Changed	What Changed	Source	Treatment Information	Outcome Measure Used	Reliability	ES	(1-_)
Outcome #1:	☐ Primary ☐ Secondary ☐ Unknown	☐ Child ☐ Teacher ☐ Parent/sign. adult ☐ Ecology ☐ Other ☐ Unknown	☐ Behavior ☐ Attitude ☐ Knowledge ☐ Other ☐ Unknown	☐ Self Report ☐ Parent Report ☐ Teacher Report ☐ Observation ☐ Test ☐ Other ☐ Unknown					
Outcome #2	☐ Primary ☐ Secondary ☐ Unknown	☐ Child ☐ Teacher ☐ Parent/sign. Adult ☐ Ecology ☐ Other ☐ Unknown	☐ Behavior ☐ Attitude ☐ Knowledge ☐ Other ☐ Unknown	☐ Self Report ☐ Parent Report ☐ Teacher Report ☐ Observation ☐ Test ☐ Other ☐ Unknown					
Outcome #3:	☐ Primary ☐ Secondary ☐ Unknown	☐ Child ☐ Teacher ☐ Parent/sign. Adult ☐ Ecology ☐ Other ☐ Unknown	☐ Behavior ☐ Attitude ☐ Knowledge ☐ Other ☐ Unknown	☐ Self Report ☐ Parent Report ☐ Teacher Report ☐ Observation ☐ Test ☐ Other ☐ Unknown					
Outcome #4:	☐ Primary ☐ Secondary ☐ Unknown	☐ Child ☐ Teacher ☐ Parent/sign. Adult ☐ Ecology ☐ Other ☐ Unknown	☐ Behavior ☐ Attitude ☐ Knowledge ☐ Other ☐ Unknown	☐ Self Report ☐ Parent Report ☐ Teacher Report ☐ Observation ☐ Test ☐ Other ☐ Unknown					
Outcome #5:	☐ Primary ☐ Secondary ☐ Unknown	☐ Child ☐ Teacher ☐ Parent/sign. Adult ☐ Ecology ☐ Other ☐ Unknown	☐ Behavior ☐ Attitude ☐ Knowledge ☐ Other ☐ Unknown	☐ Self Report ☐ Parent Report ☐ Teacher Report ☐ Observation ☐ Test ☐ Other ☐ Unknown					

Null Findings/Negative Outcomes Associated with the Intervention (listed alphabetically by outcome)

Outcomes	Primary vs. Secondary	Who Was Targeted for Change	What Was Targeted for Change	Source	Note null/negative outcomes	Outcome Measure Used	Reliability	ES
Outcome #1:	☐ Primary ☐ Secondary ☐ Unknown	☐ Child ☐ Teacher ☐ Parent/sign. Adult ☐ Ecology ☐ Other ☐ Unknown	☐ Behavior ☐ Attitude ☐ Knowledge ☐ Other ☐ Unknown	☐ Self Report ☐ Parent Report ☐ Teacher Report ☐ Observation ☐ Test ☐ Other ☐ Unknown				
Outcome #2	☐ Primary ☐ Secondary ☐ Unknown	☐ Child ☐ Teacher ☐ Parent/sign. Adult ☐ Ecology ☐ Other ☐ Unknown	☐ Behavior ☐ Attitude ☐ Knowledge ☐ Other ☐ Unknown	☐ Self Report ☐ Parent Report ☐ Teacher Report ☐ Observation ☐ Test ☐ Other ☐ Unknown				
Outcome #3:	☐ Primary ☐ Secondary ☐ Unknown	☐ Child ☐ Teacher ☐ Parent/sign. Adult ☐ Ecology ☐ Other ☐ Unknown	☐ Behavior ☐ Attitude ☐ Knowledge ☐ Other ☐ Unknown	☐ Self Report ☐ Parent Report ☐ Teacher Report ☐ Observation ☐ Test ☐ Other ☐ Unknown				
Outcome #4:	☐ Primary ☐ Secondary ☐ Unknown	☐ Child ☐ Teacher ☐ Parent/sign. Adult ☐ Ecology ☐ Other ☐ Unknown	☐ Behavior ☐ Attitude ☐ Knowledge ☐ Other ☐ Unknown	☐ Self Report ☐ Parent Report ☐ Teacher Report ☐ Observation ☐ Test ☐ Other ☐ Unknown				
Outcome #5:	☐ Primary ☐ Secondary ☐ Unknown	☐ Child ☐ Teacher ☐ Parent/sign. Adult ☐ Ecology ☐ Other ☐ Unknown	☐ Behavior ☐ Attitude ☐ Knowledge ☐ Other ☐ Unknown	☐ Self Report ☐ Parent Report ☐ Teacher Report ☐ Observation ☐ Test ☐ Other ☐ Unknown				

Type of Data Effect Size is Based On	Confidence Rating in ES Computation
(check all that apply) ☐ Means and SDs ☐ *t* - value or *F* – value ☐ Chi square (*df* = 1) ☐ Frequencies or proportions (dichotomous) ☐ Frequencies or proportions (polytomous) ☐ Other (specify): ☐ Unknown	(select one of the following) ☐ Highly estimated (e.g., only have *N* p value) ☐ Moderate estimation (e.g., have complex but complete statistics) ☐ Some estimation (e.g., unconventional statistics that require conversion) ☐ Slight estimation (e.g., use significance testing statistics rather than descriptives) ☐ No estimation (e.g., all descriptive data is present)

D. Educational/Clinical Significance

Outcome Variables:	Pretest	Posttest	Follow Up
D1. Categorical Diagnosis Data	Diagnostic information regarding inclusion into the study presented: ☐ Yes ☐ No ☐ Unknown	Positive change in diagnostic criteria from pre to posttest: ☐ Yes ☐ No ☐ Unknown	Positive change in diagnostic criteria from posttest to follow up: ☐ Yes ☐ No ☐ Unknown
D2. Outcome Assessed via continuous Variables		Positive change in percentage of participants showing clinical improvement from pre to posttest: ☐ Yes ☐ No ☐ Unknown	Positive change in percentage of participants showing clinical improvement from posttest to follow up: ☐ Yes ☐ No ☐ Unknown
D3. Subjective Evaluation: The importance of behavior change is evaluated by individuals in direct contact with the participant.	Importance of behavior change is evaluated: ☐ Yes ☐ No ☐ Unknown	Importance of behavior change from pre to posttest is evaluated positively by individuals in direct contact with the participant: ☐ Yes ☐ No ☐ Unknown	Importance of behavior change from posttest to follow up is evaluated positively by individuals in direct contact with the participant: ☐ Yes ☐ No ☐ Unknown
D4. Social Comparison: Behavior of participant at pre, post, and follow up is compared to normative data (e.g., a typical peer).	Participant's behavior is compared to normative data ☐ Yes ☐ No ☐ Unknown	Participant's behavior has improved from pre to posttest when compared to normative data: ☐ Yes ☐ No ☐ Unknown	Participant's behavior has improved from posttest to follow up when compared to normative data: ☐ Yes ☐ No ☐ Unknown

Rating for Educational/Clinical Significance (select 0, 1, 2, or 3)**: ☐3 ☐2 ☐1 ☐0**

E. Identifiable Components (answer E1 through E7)

E1. Evidence for primary outcomes (rate from previous code): ☐3 ☐2 ☐1 ☐0

~~E2. Design allows for analysis of identifiable components (select one) ☐yes ☐no~~

~~E3. Total number of components:~~

~~N~~

~~E4. Number of components linked to primary outcomes:~~

~~N~~

~~Additional criteria to code descriptively:~~

~~E5. Clear documentation of essential components (select one) yes no~~

~~E6. Procedures for adapting the intervention are described in detail (select one) ☐yes ☐no~~

~~E7. Contextual features of the intervention are documented (select one) ☐yes ☐no~~

~~**Rating for Identifiable Components** (select 0, 1, 2, or 3)**: ☐3 ☐2 ☐1 ☐0**~~

F. Implementation Fidelity

F1. Evidence of Acceptable Adherence (answer F1.1 through F1.3)

F1.1 ☐Ongoing supervision/consultation

F1.2 ☐Coding intervention sessions/lessons or procedures

F1.3 ☒Audio/video tape implementation (select F1.3.1 or F1.3.2):

F1.3.1 ☒Entire intervention

F1.3.2 ☐Part of intervention

F2. Manualization (select all that apply)

F2.1 ☐Written material involving a detailed account of the exact procedures and the sequence in which they are to be used

F2.2 ☐Formal training session that includes a detailed account of the exact procedures and the sequence in which they are to be used

F2.3 ☐Written material involving an overview of broad principles and a description of the intervention phases

F2.4 ☐Formal or informal training session involving an overview of broad principles and a description of the intervention phases

F3. Adaptation procedures are specified (select one) ☐yes ☐no ☒unknown

Rating for Implementation Fidelity (select 0, 1, 2, or 3)**: ☐3 ☒2 ☐1 ☐0**

~~**G. Replication** (answer G1, G2, G3, and G4)~~

~~G1. ☐Same Intervention~~
~~G2. ☐Same Target Problem~~
~~G3. ☐Independent evaluation~~

~~**Rating for Replication** (select 0, 1, 2, or 3)**: ☐3 ☐2 ☐1 ☐0**~~

~~**H. Site of Implementation**~~

~~H1. School (if school is the site, select one of the following options)~~

~~H1.1 ☐Public~~
~~H1.2 ☐Private~~
~~H1.3 ☐Charter~~
~~H1.4 ☐University Affiliated~~
~~H1.5 ☐Alternative~~
~~H1.6 ☐Not specified/unknown~~

~~H2. Non School Site (if it is a non school site, select one of the following options)~~
~~H2.1 ☐Home~~
~~H2.2 ☐University Clinic~~
~~H2.3 ☐Summer Program~~
~~H2.4 ☐Outpatient Hospital~~
~~H2.5 ☐Partial inpatient/day Intervention Program~~
~~H2.6 ☐Inpatient Hospital~~
~~H2.7 ☐Private Practice~~
~~H2.8 ☐Mental Health Center~~
~~H2.9 ☐Residential Treatment Facility~~
~~H2.10 ☐Other (specify):_____~~
~~H2.11 ☐Unknown/insufficient information provided~~

~~**Rating for Site of Implementation** (select 0, 1, 2, or 3)**: ☐3 ☐2 ☐1 ☐0**~~

I. Follow Up Assessment

☐Timing of follow up assessment: specify_____
☐Number of participants included in the follow up assessment: specify_____
☐Consistency of assessment method used: specify_____

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

III. Other Descriptive or Supplemental Criteria to Consider

A. External Validity Indicators

A1. Sampling procedures described in detail ☐yes ☒no
Specify rationale for selection: yes – based on students' presenting difficulties
Specify rationale for sample size: no
A1.1Inclusion/exclusion criteria specified ☐yes ☒no
A1.2 Inclusion/exclusion criteria similar to school practice ☐yes ☒no
A1.3 Specified criteria related to concern ☒yes ☐no

~~A2. Participant Characteristics Specified for Treatment and Control Group~~

~~Participants from Treatment Group~~	~~Grade/age~~	~~Gender~~	~~Ethnicity or Multiethnic~~	~~Ethnic Identity~~	~~Race(s)~~	~~Acculturation~~	~~Pri - mary Language~~	~~SES~~	~~Family Structure~~	~~Locale~~	~~Disability~~	~~Functional Descriptors~~
☐~~Child/Student~~ ☐~~Parent/caregiver~~ ☐~~Teacher~~ ☐~~School~~ ☐~~Other~~												
☐~~Child/Student~~ ☐~~Parent/caregiver~~ ☐~~Teacher~~ ☐~~School~~ ☐~~Other~~												

~~Participants from Control Group~~	~~Grade/age~~	~~Gender~~	~~Ethnicity or Multiethnic~~	~~Ethnic Identity~~	~~Race(s)~~	~~Acculturation~~	~~Pri - mary Language~~	~~SES~~	~~Family Structure~~	~~Locale~~	~~Disability~~	~~Functional Descriptors~~
☐~~Child/Student~~ ☐~~Parent/caregiver~~ ☐~~Teacher~~ ☐~~School~~ ☐~~Other~~												
☐~~Child/Student~~ ☐~~Parent/caregiver~~ ☐~~Teacher~~ ☐~~School~~ ☐~~Other~~												

A3. Details are provided regarding variables that:

A3.1 Have differential relevance for intended outcomes ☒yes ☐no

Specify: prior achievement test scores in reading, writing, maths

A3.2 Have relevance to inclusion criteria ☒yes ☐no

Specify: IQ test scores

A4. Receptivity/acceptance by target participant population (treatment group)

Participants from Treatment Group	**Results (What person reported to have gained from participation in program)**	**General Rating**
☒Child/Student ☐Parent/caregiver ☐Teacher ☐School ☐Other	Junior high participants were most often 'satisfied' with the test-taking strategy CD. They reported on the usefulness of the quizzes and gave suggestions on improving the quality of the CD presentation.	☒Participants reported benefiting overall from the intervention ☐Participants reported not benefiting overall from the intervention
☒Child/Student ☐Parent/caregiver ☐Teacher ☐School ☐Other	Senior high participants were most satisfied with the usefulness of the notes and practice activities. They were least satisfied with graphic quality.	☒Participants reported benefiting overall from the intervention ☐Participants reported not benefiting overall from the intervention
☐Child/Student ☐Parent/caregiver ☐Teacher		☐Participants reported benefiting overall from the intervention ☐Participants reported not benefiting

☐School ☐Other		overall from the intervention

A5. Generalization of Effects:

A5.1 Generalization over time

A5.1.1 Evidence is provided regarding the sustainability of outcomes after intervention is terminated ☐yes ☒no

Specify:_____

A5.1.2 Procedures for maintaining outcomes are specified ☐yes ☒no

Specify: _____

A5.2 Generalization across settings

A5.2.1 Evidence is provided regarding the extent to which outcomes are manifested in contexts that are different from the intervention context ☐yes ☒no

Specify: _____

A5.2.2 Documentation of efforts to ensure application of intervention to other settings ☐yes ☒no

Specify: _____

A5.2.3 Impact on implementers or context is sustained ☐yes ☒no

Specify: _____

A5.3 Generalization across persons

Evidence is provided regarding the degree to which outcomes are manifested with participants who are different than the original group of participants for with the intervention was evaluated ☐yes ☒no

Specify: _____

B. Length of Intervention (select B1 or B2)

B1. ☒Unknown/insufficient information provided

B2. ☐Information provided (if information is provided, specify one of the following:)

B2.1 weeks _____
N

B2.2 months _____
N

B2.3 years _____
N

B2.4 other _____
N

C. Intensity/dosage of Intervention (select C1 or C2)

C1. ☐Unknown/insufficient information provided

C2. ☒Information provided (if information is provided, specify both of the following:)

C2.1 length of intervention session 30 – 45 minutes
N

C2.2 frequency of intervention session 4 sessions
N

~~**D. Dosage Response** (select D1 or D2)~~

~~D1. ☐Unknown/insufficient information provided~~

~~D2. ☐Information provided (if information is provided, answer D2.1)~~
~~D2.1 Describe positive outcomes associated with higher dosage: _____~~

E. Program Implementer (select all that apply)

E1. ☒Research Staff
E2. ☐School Specialty Staff
E3. ☒Teachers
E4. ☐Educational Assistants
E5. ☐Parents
E6. ☐College Students
E7. ☐Peers
E8. ☒Other (CD)
E9. ☐Unknown/insufficient information provided

F. Characteristics of the Intervener

F1. ☐Highly similar to target participants on key variables (e.g., race, gender, SES)
F2. ☒Somewhat similar to target participants on key variables – CD had student actors
F3. ☐Different from target participants on key variables

G. Intervention Style or Orientation (select all that apply)

G1. ☐Behavioral
G2. ☒Cognitive-behavioral
G3. ☐Experiential
G4. ☐Humanistic/interpersonal
G5. ☐Psychodynamic/insight oriented
G6. ☐other (specify):_____
G7. ☐Unknown/insufficient information provided

H. Cost Analysis Data (select G1 or G2)
H1. ☒Unknown/insufficient information provided
H2. ☐Information provided (if information is provided, answer H2.1)
H2.1 Estimated Cost of Implementation:_____

~~**I. Training and Support Resources** (select all that apply)~~
~~I1. ☐Simple orientation given to change agents~~
~~I2. ☐Training workshops conducted~~
~~# of Workshops provided _____~~
~~Average length of training _____~~
~~Who conducted training (select all that apply)~~
~~I2.1 ☐Project Director~~
~~I2.2 ☐Graduate/project assistants~~
~~I2.3 ☐Other (please specify):~~
~~I2.3 ☐Unknown~~

~~I3. ☐Ongoing technical support~~
~~I4. ☐Program materials obtained~~
~~I5. ☐Special Facilities~~

~~I6. ☐Other (specify):~~

J. Feasibility

J1. Level of difficulty in training intervention agents (select one of the following)

J1.1 ☐High
J1.2 ☐Moderate
J1.3 ☐Low
J1.4 ☒Unknown

J2. Cost to train intervention agents (specify if known): _____

J3. Rating of cost to train intervention agents (select one of the following)

J3.1 ☐High
J3.2 ☐Moderate
J3.3 ☐Low
J3.4 ☒Unknown

Summary of Evidence for Group-Based Design Studies

Indicator	**Overall Evidence Rating** NNR = No numerical rating or 0 - 3	**Description of Evidence** Strong Promising Weak No/limited evidence or Descriptive ratings
General Characteristics		
General Design Characteristics	**NNR**	
Statistical Treatment	**NNR**	
Type of Program	**NNR**	
Stage of Program	**NNR**	
Concurrent/Historical Intervention Exposure	**NNR**	
Key Features		
Measurement	**1**	**Weak**
Comparison Group	**2**	**Promising**
~~Primary/Secondary Outcomes are Statistically Significant~~	**NA**	
~~Educational/clinical significance~~	**NA**	
~~Identifiable Components~~	**NA**	
Implementation Fidelity	**2**	**Promising**
~~Replication~~	**NA**	
~~Site of Implementation~~	**NA**	
Follow Up Assessment Conducted	**0**	**No/limited**
Descriptive or Supplemental Criteria		
External validity indicators	**NNR**	
Length of Intervention	**NNR**	
Intensity/dosage	**NNR**	
~~Dosage Response~~	**NA**	

Program Implementer	**NNR**	
Characteristics of the Intervener	**NNR**	
Intervention Style/Orientation	**NNR**	
Cost Analysis Data Provided	**NNR**	
~~Training and Support Resources~~	**NA**	
Feasibility	**NNR**	

Average WoE A across the 4 judgement areas:

Sum of X / N = 5 /4 = 1.25

X = individual quality rating for each judgement area
N = number of judgement areas

Overall rating of evidence (0-3): 1

Appendix F

Weight of Evidence A

The score for methodological quality (WoE A) is based on the rating given to each of the studies according to the coding protocol for single case designs (Horner et al., 2005) and the coding protocol for group based designs (Kratochwill, 2003).

WoE A for Single Case Designs

The WoE A criteria for single case designs is derived from Horner et al.'s (2005) paper, which describes 6 areas of professional standards that are considered in judging whether a study is of 'high' quality. In judging whether a study is of 'medium' or 'low' quality, the reviewer has drawn on criteria from both Horner et al. (2005) and Kratochwill (2003), as outlined in the below tables. Studies had to meet all of the criteria for each weighting in order to be given that judgement.

1. Description of Participants and Setting

Weighting	Criteria
High (3)	- Study must provide information on the participant recruitment process - Study meets the below criteria
Medium (2)	- Study has a clear description of participant characteristics - Study provides a detailed description of the setting in which the intervention in implemented
Low (1)	- Study either describes the setting or describes participant characteristics

2. Dependent Variable

Weighting	Criteria
High (3)	- Dependent variables generate a quantifiable index and are described in detail - Study meets the below criteria
Medium (2)	- Study uses a clearly operationalised dependent variable - Inter-rater reliability meets minimal standards (at least 0.61 Kappa or 75% agreement) - Study meets the below criteria
Low (1)	- Dependent variables are measured repeatedly over time - Information is provided on inter-rater reliability

3. Independent Variable

Weighting	Criteria
High (3)	- Fidelity of implementation is measured specifically - Study meets the below criteria
Medium (2)	- The independent variable is described in detail and manipulated in a structured way
Low (1)	- The independent variable is manipulated in a structured way

	but may not be described in sufficient detail to allow replication

4. Baseline

Weighting	Criteria
High (3)	- The dependent variable is consistently measured 3+ times during the baseline phase - Baseline conditions are described in detail
Medium (2)	- Baseline conditions are described in detail - The dependent variable is measured multiple times at baseline
Low (1)	- The dependent variable is measured multiple times at baseline

5. Experimental Control/Internal Validity

Weighting	Criteria
High (3)	- Common threats to internal validity are controlled for, to minimise the impact of extraneous variables - Study meets the below criteria
Medium (2)	- Experimental effect is demonstrated at least at 3 different points in time - Experimental control is demonstrated in the pattern of results
Low (1)	- Experimental effect is demonstrated at least at 3 different points in time

6. Social Validity

Weighting	Criteria
High (3)	- Experimental effects are found across all participants irrespective of differing characteristics - Study meets the below criteria
Medium (2)	- The dependent variable is socially important
Low (1)	- The dependent variable is not socially important

WoE A for Group-Based Designs

The WoE A criteria for group based designs is derived from Kratochwill's (2003) coding protocol, which was initially amended to suit this review's particular question (see Appendix D). The 4 key features of a study were rated on a numerical basis, according to the criteria outlined below. Studies had to meet all of the criteria for each weighting in order to be given that judgement.

1. Measurement

Weighting	Criteria
High (3)	- Study reports the type of reliability statistic used; a reliability coefficient of 0.70 or higher is found for the majority of outcomes - Data is collected from multiple methods and sources - Validity of measures is reported

Medium (2)	- Data is collected from multiple methods and/or multiple sources - Validity of measures is reported
Low (1)	- A case for validity does not need to have been presented - Data is collected from multiple methods or multiple sources

2. Comparison Group

Weighting	Criteria
High (3)	- Study must explicitly describe the type of comparison group - There should be at least one 'active' comparison group - Group equivalence must be demonstrated statistically - Low attrition must be reported
Medium (2)	- Study must use at least a 'no intervention' comparison group - At least two of the following must also be present: counterbalancing of change agents, group equivalence or low attrition
Low (1)	- Study must use a comparison group - At least one of the following must also be present: counterbalancing of change agents, group equivalence or low attrition

3. Implementation Fidelity

Weighting	Criteria
High (3)	- Study demonstrates strong evidence of acceptable adherence to implementation fidelity - Evidence should involve either ongoing supervision or the coding of intervention sessions - A detailed manual with the exact procedures and sequence should be followed
Medium (2)	- Study demonstrates evidence of acceptable adherence to implementation fidelity - Evidence involves audio/videotape implementation for all of the intervention
Low (1)	- Study demonstrates some evidence of acceptable adherence, through at least one of the above criteria, or the use of a manual

4. Follow Up Assessment

Weighting	Criteria
High (3)	- Study conducts follow up assessments over multiple intervals with the majority of participants that were included in the original sample - There is consistency in the assessment methods used
Medium (2)	- Study conducts follow up assessments at least once; the timing of which is specified - Follow up assessment includes some of the participants that were included in the original sample - There is consistency in the assessment methods used
Low (1)	- Study conducts one follow up assessment - Follow up assessment includes some of the participants

	included in the original sample - A case is not made for consistency in the assessment methods used

Each of the dimensions (1-6 for single case designs and 1-4 for group designs) was rated numerically from 0-3 with a score of 0 indicating 'no evidence', 1 indicating 'weak evidence', 2 indicating 'promising evidence' and 3 indicating 'strong evidence'. The weightings from each section were then averaged to give an overall measure of methodological quality (WoE A).

To receive a 'high' rating overall for methodological quality, the study must receive an average rating of 2.5 or above. To receive a 'medium' rating overall for methodological quality, a study must receive an average rating between 1.5 and 2.4. To receive a 'low' weighting overall for methodological quality, the study must receive an average rating of 1.4 or below.

Summary of overall WoE A ratings across studies

Single Case Studies

Authors	**Dimensions**						**Overall WoE A**
	Participants & Setting	*Dependent Variable*	*Independent Variable*	*Baseline*	*Experimental Control*	*Social Validity*	
Hughes & Schumaker (1991)	2	3	3	2	2	3	**High** (2.5)
Lancaster et al. (2006)	2	2	2	3	2	3	**Medium** (2.3)

Group Based Studies

Authors	**Dimensions**				**Overall WoE A**
	Measurement	*Comparison Group*	*Implementation Fidelity*	*Follow Up*	
Carter et al. (2005)	2	3	3	2	**High** (2.5)
Lancaster et al. (2009)	1	2	2	0	**Low** (1.25)
Therrien et al. (2009)	1	3	3	0	**Medium** (1.75)

Weight of Evidence B

The score for methodological relevance (WoE B) is based on each study's design appropriateness, given the specific review question. In other words, WoE B considers whether the methodological design is suitable for evaluating the effectiveness of test-taking strategy interventions on young people with learning difficulties.

The criteria for WoE B are based on evidence hierarchies (Guyatt et al., 2008). These hierarchies typically place studies with minimal threats to internal validity at the top, whilst those with high threats to internal validity and single case designs are generally placed lower down. The additional criterion of pre, post and follow up measures is in place because maintenance of skills is particularly relevant for the review's population of interest, therefore a study that demonstrates lasting effects of their intervention will be given more weight. Studies had to meet all of the criteria for each weighting in order to be given that judgement.

Weighting	Criteria
High (3)	- Study must have a group design with an 'active' comparison group - Outcome data must be collected at pre, post and follow up - Study makes use of randomisation
Medium (2)	- Single case studies must provide at least 3 demonstrations of experimental effect - Group based studies have a control group and outcome data is collected pre and post intervention for both groups
Low (1)	- Single case design studies may not have demonstrated intervention effect 3 times. - Group based designs may not have a control group and outcome data is collected at pre and post intervention

WoE B ratings across studies

	Authors	Methodological Relevance WoE B
1	Carter, Wehby, Hughes, Johnson, Plank, Barton-Arwood & Lunsford (2005)	**Medium** (2)
2	Hughes & Schumaker (1991)	**Medium** (2)
3	Lancaster, Lancaster, Schumaker & Deshler (2006)	**Medium** (2)
4	Lancaster, Schumaker, Lancaster & Deshler (2009)	**Medium** (2)
5	Therrien, Hughes, Kapelski & Mokhtari (2009)	**Medium** (2)

Weight of Evidence C

The score for topic relevance (WoE C) is based on the appropriateness of each study's focus to the specific review question. The judgements were made based on the following rationale.

- The extent to which the test-taking strategy intervention is described in the study is considered important criteria. This should include details of the mnemonic device, particularly what the expectation is of the test-taker at each step of the strategy. Some evidence of the 8 procedural steps outlined by Mercer and Pullen (2005) in table 2 would also be favourable, as would specific attention being given to fidelity of implementation. Taken together, these would help to ensure that the intervention is indeed tapping in to the cognitive and metacognitive skills that the target population have difficulty with.
- The sample should also be central to the target population of the review, i.e. secondary aged students identified as having learning difficulties. If the sample also have other presenting difficulties, such as additional emotional or social needs, this may impact on the generalisation of the findings to the target population.
- Outcome measures that focus on academic attainment, particularly those in relation to summative or formative assessment at school, would be the most relevant to the review question as they would give an indication of ecological validity. Additionally, Multi-sourced or multi-method measurement would show evidence of triangulating data, which provides a richer report of the outcomes of the intervention.

Weighting	Criteria
High (3)	Study meets all of the below: - The study explains the test-taking strategy in depth, including information around the procedural steps - The sample consists only of secondary aged students with learning difficulties - Usual testing used at the participants' schools are utilised as part of the academic attainment outcomes. - There is evidence of triangulation of data
Medium (2)	Study meets 3 of the below: - The study explains the test-taking strategy in depth, including information around the procedural steps - The sample consists only of secondary aged students with learning difficulties - Usual testing used at the participants' schools are utilised as part of the academic outcomes - There is evidence of triangulation of data
Low (1)	Study meets 1 or 2 of the below: - The study explains the test-taking strategy in depth, including information around the procedural steps - The sample consists only of secondary aged students with learning difficulties - Usual testing used at the participants' schools are utilised as part of the academic outcomes - There is evidence of triangulation of data

Note: if the criteria for a 'low' weighting are not met, a score of 0 is awarded.

WoE C ratings across studies

	Authors	Topic Relevance WoE C
1	Carter, Wehby, Hughes, Johnson, Plank, Barton-Arwood & Lunsford (2005)	**Low** (1)
2	Hughes & Schumaker (1991)	**Medium** (2)
3	Lancaster, Lancaster, Schumaker & Deshler (2006)	**Medium** (2)
4	Lancaster, Schumaker, Lancaster & Deshler (2009)	**Medium** (2)
5	Therrien, Hughes, Kapelski & Mokhtari (2009)	**Medium** (2)

Weight of Evidence D

Using the criteria outlined above, each study was given a weighting for WoE A, B and C as either 3 'High', 2 Medium' or 1 'Low'. These scores were then averaged to provide each study with an overall Weight of Evidence score (WoE D). Essentially, WoE D considers the extent to which the study, and its findings contribute towards answering the question, 'how effective are test-taking strategy interventions in improving the academic attainment of secondary aged students with learning difficulties?'

To receive an overall weighting of **'High'**, the study must receive an average score of 2.5 or above. To receive an overall weighting of **'Medium'**, the study must receive an average score of between 1.5 and 2.4. To receive an overall weighting of **'Low'**, the study must have an average score of 1.4 or below.

Overall WoE D across studies

Authors	WoE A: Quality of Methodology	WoE B: Relevance of Methodology	WoE C: Study Topic Relevance	WoE D: Overall Weight of Evidence
Carter et al. (2005)	High (2.5)	Medium (2)	Low (1)	**Medium** (1.83)
Hughes & Schumaker (1991)	High (2.5)	Medium (2)	Medium (2)	**Medium** (2.17)
Lancaster et al. (2006)	Medium (2.3)	Medium (2)	Medium (2)	**Medium** (2.1)
Lancaster et al. (2009)	Low (1.25)	Medium (2)	Medium (2)	**Medium** (1.75)
Therrien et al. (2009)	Medium (1.75)	Medium (2)	Medium (2)	**Medium** (1.91)