Case Study 1: An Evidence-Based Practice Review Report
Theme: School Based Interventions for Learning

Is Solve it! effective in improving the mathematics achievement of students with mathematics difficulties?

Summary
Cognitive strategy instruction comprises instruction in cognitive and metacognitive processes. It has been shown to be effective in meeting the learning needs of students with Learning Disabilities. This systematic literature review aims to evaluate the effectiveness of a cognitive strategy intervention – Solve it! – in improving the mathematics attainment of students with mathematics difficulties. Four group experimental design and two single-case-design studies were critically appraised. Solve it! is found to be effective in improving mathematical word problem solving skills. However, limited evidence exists to support the generalisability and sustainability of the intervention’s effects. Limitations identified in the empirical literature and suggestions for future research are discussed.
Introduction

_Solve it!_ is an intervention that aims to teach students to be good problem solvers (Montague, 2003). It is a “comprehensive strategic routine consisting of seven cognitive processes (read, paraphrase, visualise, hypothesise, estimate, compute, and check) and corresponding self-regulation strategies (self-instruction, self-questioning, and self-monitoring)” (Montague, Enders & Dietz, 2011, p. 263). It was originally developed by Montague (2003) to improve the mathematical word problem solving of students with Learning Disabilities (LD).

_Solve it!_ involves a scripted curriculum where teachers explicitly teach and guide students through a series of steps to solve word problems. The overall structure of the intervention involves two main parts. First students are taught the problem solving routine through scripted lessons which incorporate explicit instructional procedures. After learning the routine, students then engage in word problem solving exercises where they practice the strategy. Students’ use of the strategy and performance are monitored throughout the intervention.

All the materials mentioned above and others (Math Problem Solving Assessment, instructional charts, activities, cue cards) are provided in Montague (2003). It has been implemented in whole-class, small group and individual settings. Among the papers in this review, students were instructed individually in the Single Case Design (SCD) studies, and in a whole-class setting for group studies.

**Theoretical Underpinnings**

Three key features of _Solve it!_ are rooted in psychological theory. Firstly, the instructional approach is based on cognitive strategy instruction. Next, the instructional content is based on problem-solving models. Lastly, the procedural
basis is based on explicit instruction.

*Solve it!*’s instructional approach is based on cognitive strategy instruction, which comprises instruction in cognitive processes (e.g. paraphrasing) and metacognitive/self-regulation strategies (e.g. self-instruction) (Montague & Dietz, 2009; Wong, Harris, Graham & Butler, 2003). “Metacognition refers to one’s knowledge concerning processes and products… to the active monitoring and consequent regulation and orchestration of these (thinking) processes” (Flavell, 1976, p.232). It has been established that learning is supported by metacognition (Dole & Sinatra, 1998). Wong et al. (2003) provides a comprehensive review of cognitive strategy instruction research.

While the instructional approach is based on cognitive strategy instruction, the instructional content is based on problem-solving models. The seven cognitive processes that comprise the intervention are based on traditional problem-solving models (Mayer, 1985; Polya, 1986). These models are well-supported by research and also align with current conceptualisations of mathematical problem solving processes (OECD, 2013). Polya (1986) described four strategies: understand the problem; devise a plan; carry out the plan; look back. *Solve it!*’s seven cognitive processes align with Polya’s model: understand the problem (read, paraphrase, visualise), devise a plan (hypothesise, estimate), carry out the plan (compute), look back (check). The cognitive processes also align with the OECD framework (2013): exploring and understanding (read, paraphrase), representing and formulating (visualise), planning and executing (hypothesise, estimate, compute), and monitoring and reflecting (check).

Finally, explicit instruction is the procedural approach of *Solve it!* “Explicit instruction is characterized by highly structured and organised lessons, appropriate
cues and prompts, guided and distributed practice, cognitive modelling, interaction between teachers and students, immediate and corrective feedback on performance, positive reinforcement, overlearning and mastery” (Montague, 2003). Research has established that explicit instruction is an effective intervention for students with mathematics difficulties (Baker, Gersten & Lee, 2002; Fuchs, Fuchs, Schumacher & Seethaler, 2013; Kroesberg & Van Luit, 2003). The use of cognitive modelling – thinking aloud while modelling a cognitive strategy – is critical to cognitive strategy instruction as it allows students to see how effective problem solvers behave and consequently, to learn by imitation (Montague & Dietz, 2009).

Thus, the theoretical foundation of Solve it! is rooted in developmental, information processing, and sociocultural theories. The use of explicit instruction by the teacher reflects Vygotsky’s (1978) social development theory, while the use of cognitive processes aligns with the information processing view of specifying the knowledge that a learner needs to learn to solve an academic task (Mayer, 2003).

**Rationale and Relevance to EP Practice**

*Mathematical problem solving*

Over the last two decades, efforts to improve mathematics achievement have directed their attention onto word problem solving (Carr, 2012; MacArthur, 2012). Word problem solving is recognised as an important medium through which children learn strategies for applying their knowledge of mathematics to solving problems in the real world (Swanson, 2014). Mathematical problem solving is one of the most difficult components of the mathematics curriculum and success in mathematical word problem solving is highly correlated with overall mathematics achievement (Bryant, Bryant & Hammill, 2000; Hudson & Miller, 2006). Moreover, studies have shown that the cognitive processes involved in word problem solving are different
from those involved in computation (Fuchs et al., 2008). Thus, interventions specifically addressing word problem solving difficulties are necessary.

*Cognitive strategy instruction and students with LD*

A comprehensive review of cognitive strategy instruction research by Wong et al. (2003) shows that cognitive strategy instruction is particularly effective for students with LD. For example, self-regulated strategy development (SRSD) has been established as an effective writing intervention and is based on cognitive strategy instruction (for a more detailed discussion of SRSD, see Harris & Graham, 1999).

Students with LD often use ineffective strategies and do not generalise strategy use across learning (Swanson, 1993). Conversely, strategic learners can effectively and efficiently adapt their strategies according to task demands (Pressley, Borkowski, & Schneider, 1987). Students with LD struggle especially with mathematical word problem solving as it is a complex process involving many steps (Fuchs et al., 2008, Swanson, 2006). *Solve it!* is well-suited for students with LD as it explicitly teaches them the cognitive and metacognitive tools they need to solve mathematical word problems.

*The important role of metacognition*

Metacognition is especially crucial to mathematical word problem solving as it requires regulating, monitoring and evaluating performance (Mayer, 1998). It is important for mathematics remediation to incorporate metacognitive strategies as students with LD often display attention, motivation and self-regulation difficulties that may adversely affect their learning (Fuchs et al., 2013). Word problem solving strategies that combine cognitive and metacognitive procedures produce more positive effects than interventions that only utilise a single strategy (Jitendra et al.,
2015). *Solve it!* comprises both cognitive and metacognitive strategies, and aims to develop strategic learners who are also motivated, self-directed and self-regulated, and who have an effective and efficient repertoire of strategies.

**Relevance to EP practice**

Supporting children and young people with special educational needs and disabilities is well within the remit of the EP, and this includes those who are struggling with their learning. As mentioned earlier, students with learning difficulties tend to find mathematical problem solving difficult. EPs are often asked to recommend strategies to raise attainment not only at the individual level, but also at the whole-school level. The interventions that EPs recommend to school decision-makers must be evidence-based and this systematic review hopes to inform EPs of whether *Solve it!* is an evidence-based strategy to improve mathematics achievement. The last systematic review by Montague and Dietz (2009) found insufficient evidence to support cognitive strategy instruction as an evidence-based practice. Nevertheless, recent studies have found cognitive strategy instruction effective in improving mathematical word problem solving (Montague, Enders, & Dietz, 2011). The review question asked is: Is *Solve it!* effective in improving the mathematics achievement of students with mathematics difficulties?

**Method**

A literature search of the following online databases – PsychINFO, ERIC, British Education Index (BEI), MEDLINE and Web of Science – was carried out using the search terms ['cognitive strateg* instruction' OR 'strateg* instruction' or 'cognitive strategy*'] AND mathematics.

A filter was imposed on the online databases to only include studies which

---

1 /" represents wildcard search item.
were published after 2006, written in the English language and from a peer reviewed journal (see Table 1 for rationale). The search generated 194 results in total, of which 79 were duplicates and excluded. The remaining 115 abstracts were reviewed according to the inclusion criteria (Table 1), which resulted in the identification of 10 studies. These 10 articles were read in full to see if they adhered to the inclusion criteria and, consequently, 4 articles were excluded. Figure 1 provides a visual representation of this process. Table 2 lists the studies reviewed (see Appendix A for detailed summaries). Studies that were excluded after full-text reading are presented in Appendix B.

Table 1. Inclusion and Exclusion Criteria

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Inclusion</th>
<th>Exclusion</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Type of Publication</td>
<td>Study is published in a peer-reviewed journal</td>
<td>Study is not published in a peer-reviewed journal</td>
<td>The study has a level of methodological rigour from a process of external moderation.</td>
</tr>
<tr>
<td>2 Language</td>
<td>Study is written in English</td>
<td>Study is not written in English</td>
<td>Translation is not feasible within the time and cost constraints of Case Study 1.</td>
</tr>
<tr>
<td>3 Year of Publication</td>
<td>Study is published between 2007 and 2015 inclusive.</td>
<td>Study is published before 2007.</td>
<td>Montague &amp; Dietz (2009) last conducted a review on cognitive strategy instruction, which looked at papers published up to 2006. The 2015 cut-off is to allow for results to be collated to meet the case study 1 deadline.</td>
</tr>
<tr>
<td>Criteria</td>
<td>Inclusion</td>
<td>Exclusion</td>
<td>Rationale</td>
</tr>
<tr>
<td>----------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td>4 Type of study</td>
<td>An empirical study which uses a research design that is experimental, quasi-experimental or SCD.</td>
<td>The study is not empirical e.g. report, review, book chapter.</td>
<td>This ensures that primary data is being reviewed.</td>
</tr>
<tr>
<td>5 Intervention</td>
<td>The intervention was <em>Solve it!</em> If not explicitly stated, the intervention met the definition of cognitive strategy instruction, as defined in the introduction</td>
<td>Interventions that were strategy instruction but did not comprise metacognitive or self-regulation component(s).</td>
<td>This relates directly to the review question which is to evaluate <em>Solve it!</em></td>
</tr>
<tr>
<td>6 Participants and Setting</td>
<td>The study included students with a disability, as defined by the authors; and the study was conducted in school.</td>
<td>Studies that did not include students with disabilities e.g. studies that were conducted with typically developing students; or were not conducted in school.</td>
<td>This links directly to the review question and to the overall theme of school-based interventions.</td>
</tr>
<tr>
<td>7 Outcome Measure</td>
<td>Mathematical problem solving was an outcome measure.</td>
<td>Studies that do not report math problem solving performance.</td>
<td>This ensures that papers relate directly to the research question.</td>
</tr>
</tbody>
</table>

*Table 2. Studies included in this review*

**Included Studies**


Included Studies


Critical Review of the Evidence Base

The six studies were evaluated using the Weight of Evidence (WoE) framework (Gough, 2007). The WoE framework allows for studies to be evaluated on the quality of methodology (WoE A), relevance of methodology to the review (WoE B), relevance of study focus to the review (WoE C), and an overall judgment of the extent of the study’s contribution to the review (WoE D). For WoE A, group studies (Babakhani, 2011; Montague et al., 2011; 2014; Zhu, 2015) were evaluated using the quality indicators for group studies (Gersten et al., 2005) while SCD studies (Coughlin & Montague, 2010; Whitby, 2012) were evaluated using the quality indicators within single-subject research (Horner et al., 2005). Table 3 summarises the WoE ratings for each study in this review. Details of all ratings can be found in Appendix C.

Table 3. Summary of Weight of Evidence Ratings

<table>
<thead>
<tr>
<th>Study</th>
<th>Quality of methodology (A)</th>
<th>Relevance of methodology (B)</th>
<th>Relevance of evidence to the review question (C)</th>
<th>Overall weight of evidence (D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Babakhani (2011)</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Montague et al. (2014)</td>
<td>High</td>
<td>High</td>
<td>Med</td>
<td>High</td>
</tr>
</tbody>
</table>

Participants
**Demographic characteristics**

The number of participants in the studies ranged from 3 – 1059 students. Participants ranged in age from 9 to 17 years, encompassing both primary and secondary aged students. Participants were students from a range of abilities – average achievers; low achievers; students with learning disabilities, Autism Spectrum Disorder (ASD) or Spina Bifida. Four studies were conducted in the United States (Coughlin & Montague, 2010; Montague et al., 2011; 2014; Whitby, 2012), one in China (Zhu, 2015) and one was unknown (Babakhani, 2011).

Gender was balanced across most of the group studies. However, the treatment group in Zhu’s (2015) study had a slight gender imbalance (63% males, 35% females). Zhu also did not explicitly report the age of participants, although it was reported that participants were fourth-grade students. One SCD study (Whitby, 2011) did not have any female participants. This is perhaps reflective of the gender imbalance in the population diagnosed with ASD (Rice, 2009), which was the disability investigated in the study.

The level of detail provided by researchers varied when it came to demographic characteristics such as ethnicity and socio-economic status. Among the group studies, Montague et al. (2011; 2014) provided this information, but this was not reported in Babakhani (2011) and Zhu (2015). The SCD studies provided more comprehensive demographic information.

**Definition of math difficulties**

There was an inconsistency in the definition of math disability/difficulty across the studies, which is the population of focus in this review. This could be symptomatic of the current inconsistency in the diagnosis of math disability/difficulty (Gersten, Clarke, & Mazzocco, 2007; Mazzocco, 2007). Nevertheless, the current
consensus in research is that children who score at or below the 10th percentile on standardised mathematics achievement tests are categorised as having math disability, while those scoring between 11th and 25th percentiles are considered at risk of math difficulty (Geary, 2013). Gersten et al. (2005) contend that researchers should provide a definition of the disability/difficulty being studied and then include assessment results to demonstrate that participants met the definition for inclusion. They further suggest that researchers link their operational definitions to definitions in current literature where possible. The only study (Babakhani, 2011) that screened participants for inclusion used an operational criteria of scores ‘between the average and one standard deviation less than average’. The other three group studies relied on district and municipality information.

The two SCD studies provided information to demonstrate that participants met definitions for inclusion, which contributes favourably to WoE A. However, the populations investigated in their studies are not the focus in this review. As the review question is interested in children with mathematics difficulties, this consideration was reflected in WoE C, where none of the studies received a ‘high’ rating.

**Group equivalence**

Group equivalence was demonstrated in all of the group studies except Babakhani (2011). Although it was reported that participants were matched on intelligence, age and math performance, no statistical analysis of group equivalence was performed. This contributed to the study’s ‘low’ WoE A rating.

**Design**

All four group studies utilised a randomised block design. A noteworthy feature across all the group studies was the use of appropriate sample sizes. A
power analysis that was conducted indicated that a minimum sample size of 26 participants per group is required in order to detect a large effect, for tests of 0.80 power. A large effect size was chosen as a recent meta-analysis of interventions for students with mathematical difficulties found a large effect (Chodura, Kuhn, & Holling, 2015). The use of an appropriate sample size influences the generalisability of the findings, and is thus reflected in the WoE B criteria. The two SCD studies did not receive a ‘high’ rating on WoE B as the small sample sizes limited the generalisability of the findings.

**Intervention**

Four of the six studies implemented the *Solve it!* curriculum as described earlier, i.e. instruction comprised seven cognitive and three metacognitive processes. However, while the intervention in Babakhani (2011) involved all seven cognitive processes, it only utilised one of the metacognitive processes – the self-instruction ‘SAY’ procedure. The intervention in Coughlin and Montague (2010) was modified to remove the estimation procedure after the pilot study. The researchers provided the following justification – ‘estimation requires flexible thinking, which has been identified as a weakness for children with Spina Bifida [the participants of interest in their study] (Snow, 1999)’. While it would have been of interest to analyse the impact of these differences on problem solving, this was not the focus of the current review.

Implementation of the intervention varied in length and frequency. Length of intervention ranged from eight weeks to one academic year. The two Montague et al. studies (2011; 2014) started with three days of instruction and continued with weekly problem solving sessions for the rest of the academic year. The intervention conducted in Zhu (2015) entailed two 40-min lessons weekly for eight weeks, a total
of 16 sessions. It was unclear how the study was implemented in Babakhani (2011) – it was reported that implementation took place in “16 weeks of 45 minutes during two months of school hours”. The participants in the SCD studies received 8 – 10 sessions of intervention.

**Treatment Fidelity**

The two SCD studies utilised a multiple baseline across participants design. Only one (Whitby, 2012) reported measurement of treatment fidelity. Out of the four group studies, three reported treatment fidelity. As Solve it! provides scripted lessons and a treatment-fidelity checklist, fidelity of implementation was expectedly high: an average of 90% and above. Regardless of these safeguards, the studies that did not have a formal measure of treatment fidelity (Babakhani, 2011; Coughlin & Montague, 2010) did not meet the criterion within WoE A.

**Interventionists**

The interventionists in each of the studies also varied. The intervention was implemented by the researcher in three studies (Babakhani, 2011; Coughlin & Montague, 2010; Whitby, 2012) and implemented by teachers in the other three studies. Information about interventionists is an essential quality indicator for group studies (Gersten et al., 2005) and this information was specified in three of the group studies. For example, Montague et al. (2011; 2014) specified that teachers had to be “high-quality,” certified to teach math, and teach at least 2-3 class periods that included students who were low-achieving and identified as having a learning disability. In both Montague et al. studies (2011 and 2014), intervention teachers received a three day training prior to conducting the intervention, while intervention teachers in Zhu (2015) received a two day training. There was no background information about the researcher who was the interventionist in Babakhani (2011)
and this study consequently received a ‘low’ WoE A rating.

**Outcome measures**

All six studies used word problem solving as an outcome measure. Four studies used curriculum-based measures (CBM) which were based on items in the intervention manual. Both Babakhani (2011) and Zhu (2015) created their own word problem solving measures. The reliability of measures used in both studies was above 0.70. Kratochwill (2003) suggests that measures should have a reliability of at least 0.70, and this criteria is reflected in WoE B.

Zhu (2015) reported that mathematical problems tested in the outcome measure were not the same as the mathematical problems used for instruction, however this was not reported in Babakhani. In the group studies by Montague (2011; 2014), the CBM was also administered more often to the intervention group (seven times) than to the control group (four times). Montague et al. explained that this was to account for missing data and because “teachers and students might view seven administrations as excessive given that they were not receiving any direct benefit and no feedback” (Montague et al., 2014, p. 474). However, this remains a threat to internal validity due to concerns about practice effects.

**Generalisation**

According to Gersten et al. (2005), a combination of both aligned measures and more general measures (e.g. a standardised math problem solving test) are desirable. Both Montague et al (2011) and Whitby (2012) administered one Florida Comprehensive Assessment Test (FCAT) word problem, while Montague et al. (2014) collected data on participants’ FCAT Math exam scores. FCAT is the standardised test used in Florida schools, where the studies were conducted.

**Maintenance/Follow-up**
None of the four group studies administered a follow-up measure to capture maintenance of performance. On the other hand, both SCD studies administered follow-up measures to assess maintenance approximately one month later. These concerns – of generalisability and maintenance – were reflected in the WoE C ratings, and none of the studies received a ‘high’ rating.

**Findings**

Only Montague et al. (2014) and Zhu (2015) reported effect sizes. Effect sizes for group studies were calculated using Cohen’s $d$ and effect sizes for SCD studies were calculated using Percentage of Non-overlapping Data (PND). For ease of comparison, Table 4 shows the different effect sizes and their corresponding descriptors (Cohen, 1988; Scruggs & Mastropieri, 1998). Table 5 presents the effect sizes and overall WoE for each study.

Table 4. *Type of effect size and corresponding descriptors*

<table>
<thead>
<tr>
<th>Descriptors</th>
<th>Small</th>
<th>Medium</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohen’s $d$ (1988)</td>
<td>0.2</td>
<td>0.5</td>
<td>0.8</td>
</tr>
<tr>
<td>PND (Scruggs &amp; Mastropieri, 1998)</td>
<td>50-70%</td>
<td>71-90%</td>
<td>&gt;90%</td>
</tr>
</tbody>
</table>

Table 5. *Effect size and overall WoE for each study*

<table>
<thead>
<tr>
<th>Study</th>
<th>Sample size</th>
<th>Measure</th>
<th>Effect size</th>
<th>Effect size descriptor</th>
<th>Overall WoE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Babakhani (2011)</td>
<td>60</td>
<td>Verbal Math Problem Solving</td>
<td>$d = 0.49$</td>
<td>Low-Med</td>
<td>Low</td>
</tr>
<tr>
<td>Montague et al. (2011)</td>
<td>779</td>
<td>CBM</td>
<td>$d = 0.77^2$</td>
<td>Med-Large</td>
<td>Medium</td>
</tr>
<tr>
<td>Montague et al. (2014)</td>
<td>1059</td>
<td>CBM</td>
<td>$d = 0.61$</td>
<td>Medium</td>
<td>High</td>
</tr>
<tr>
<td>Zhu (2015)</td>
<td>150</td>
<td>Mathematical word problems</td>
<td>$d = 1.85$</td>
<td>Large (MD only)</td>
<td>Medium</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$d = 1.64$</td>
<td>(MD/RD)</td>
<td></td>
</tr>
</tbody>
</table>

$^2$ Effect size was calculated with data provided through correspondence with the author.
Overall, the studies indicated that the intervention had a medium to large effect. However, these findings need to be interpreted alongside the studies’ overall WoE. For example, although Zhu (2015) found a large effect size, its overall WoE is medium, and thus should be interpreted with caution. Zhu’s study included students with mathematics disability, and students with both mathematics and reading disabilities. They found the intervention had a larger effect on students with mathematics disabilities only (d = 1.85) compared to those with mathematics and reading disabilities (d = 1.64).

Montague et al.’s (2014) finding of a medium effect size is worth noting, as the study had a high overall WoE. Furthermore, the study reported a large effect size (d = .882), based on data synthesised across both 2011 and 2014 studies. This suggests that the intervention has a medium to large effect size on mathematics word problem solving skills.

All the participants in the SCD studies benefitted from the intervention, with medium to large effect sizes. It is worth noting that one participant in Coughlin and Montague’s study (2010) reported that the intervention had improved his self-esteem. While it is intuitive to investigate the impact of the intervention on mathematics, it could also be worthwhile looking into the intervention’s impact on motivation and self-regulation.
Generalisation

Both Montague et al. (2011; 2014) studies found insufficient evidence to support an intervention effect on FCAT math scores. A related issue is the presence of high attrition (>30%) in both these studies: it was found that FCAT test scores were predictive of missing data and was thus included as a covariate in the statistical analyses. Two types of FCAT math scores were collected in the study: FCAT test scores from the state-wide assessment, as well as a single FCAT word problem which was administered as an outcome measure as part of the study. In the Montague (2011) study, once FCAT test scores from the state-wide assessment were included as a covariate, differences in performance on the single FCAT word problem between treatment and control groups were insignificant.

Whitby (2012) administered a generalisation probe – one FCAT math word problem to students in the participants’ general education classroom. This was administered once in each phase. Although the data suggests that there is a positive intervention effect, there are insufficient data points to make reliable conclusions. Hence, there is a lack of evidence to support the intervention’s impact on overall mathematics achievement.

Conclusions and Recommendations

The purpose of this review is to evaluate the effectiveness of Solve it! in improving the mathematics attainment of students with mathematics difficulties. Based on the six studies evaluated, Solve it! was effective in improving students’ mathematics word problem solving skills. This is an encouraging finding, given the previous Montague and Dietz (2009) review did not find sufficient evidence to support cognitive strategy instruction as evidence-based practice. This review also supports existing literature (Maccini, Mulcahy & Wilson, 2007; Swanson, 1999) that
cognitive strategy instruction is an effective intervention in supporting students with difficulties.

As mentioned earlier, mathematical problem solving is one of the most difficult components of the mathematics curriculum and success in mathematical word problem solving is highly correlated with overall mathematics achievement (Bryant et al. 2000; Hudson & Miller, 2006). This suggests a need for effective problem solving interventions that can improve math performance. EPs are in a position to make recommendations of evidence-based interventions to school leaders and school staff so as to narrow the attainment gap. As Solve it! can be readily applied in diverse instructional contexts (e.g. at individual, classroom or whole-school levels), it is a useful strategy that EPs can recommend to alleviate students’ math problem solving difficulties. EPs can also train school staff in the strategy so that teachers are better equipped to improve students’ math performance. Nevertheless, there were some limitations to the findings and recommendations for further research are discussed below.

Limitations and Recommendations

Generalisability

While there were positive effects of the intervention on mathematics word problem solving, the outcome measures were closely aligned to the intervention. There was also limited evidence to support the intervention’s effect on general mathematics attainment. Evidence of a positive outcome on a standardised mathematics measure would lend support to the impact of the intervention. Students with learning disabilities tend to have difficulties with generalising what they learn in one domain to other aspects of learning (Montague, 2008). Further research into the impact of the intervention on overall math attainment is thus crucial.
Sustainability

While the researchers reported positive outcomes of mathematics word problem solving, there was no evidence of the sustainability of these effects. A follow-up assessment could be conducted to reinforce the findings.

Other aspects of learning and behaviour

It would be worthwhile to examine the effect of Solve it! on other areas, e.g., motivation, self-esteem, and reading comprehension. For example, Montague (2003) stated that one of Solve it!’s aims is to improve reading comprehension. A report from one participant in the SCD study of how the intervention helped to improve his self-esteem was also noted, and worth exploring systematically.

Differential impact by age

While the findings were positive overall, it would be worth investigating whether the intervention causes a differential impact on primary and secondary school students. Initial studies by Montague (1992) found that sixth-grade students with LD had difficulties accessing the instruction as they may not be maturationally ready. This could be related to the developmental trajectory of metacognitive development in children.

Investigation into processes and mechanisms

A unique feature of Solve it! is the combination of metacognitive/self-regulation strategies with instruction in a mathematics problem-solving strategy. While there is evidence for a positive effect of the intervention, further research can look into the mechanisms of how and why Solve it! is successful. We know that struggling students have difficulties monitoring the effectiveness of their problem-solving strategy use, but we lack knowledge of how to remediate it (Carr, 2003). It will thus be worthwhile investigating the impact of the intervention on problem solving
processes and metacognitive/self-regulation skills. This could be done through multi-method and multi-source approaches, e.g. student report and observations. The use of multi-method and multi-source approaches would also help produce more robust findings.

Definition of mathematics difficulties

Finally, there is a lack of consistency in defining students with mathematics difficulties. In line with Gersten et al. (2005), further studies into interventions for students with mathematics difficulties should provide sufficient information to determine that participants demonstrate mathematics difficulties. It is also recommended that researchers attempt to link their operational definitions to those in current literature, as it would help in determining the generalisability of study findings.
References


## Appendix A: Mapping the Field

<table>
<thead>
<tr>
<th>Study</th>
<th>Sample</th>
<th>Intervention</th>
<th>Measures</th>
<th>Main Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Babakhani (2011)</td>
<td>2 groups of 30 students each</td>
<td>Design: Randomised block design by gender</td>
<td>Verbal math problem solving (VMPS) test: 10 verbal math problems selected from the fourth-grade math textbook. Two measures (Form A and B) – administered at pre and post-intervention.</td>
<td>The intervention group performed significantly better on VMPS test, compared to control group. The effect was similar for boys and girls.</td>
</tr>
<tr>
<td></td>
<td>Gender: Male (n=30), Female (n=30)</td>
<td>Intervention: modified <em>Solve it!</em> of 7 cognitive processes and 1 metacognitive process (self-instruction SAY procedure). Implemented over “16 weeks of 45 minutes during two months of school hours”.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age: 9.5 – 10.5 years</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Grade: 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ethnicities, free/reduced lunch and setting: not reported</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coughlin &amp; Montague (2010)</td>
<td>3 adolescents diagnosed with Spina Bifida</td>
<td>Design: Multiple baseline across participants</td>
<td>Curriculum Based Measures (CBM): word problems taken from the <em>Solve it!</em> manual.</td>
<td>All participants improved in math word problem solving and maintained performance 1 month following instruction.</td>
</tr>
<tr>
<td></td>
<td>Gender: Male (n=2), Female (n=1)</td>
<td>Intervention: modified <em>Solve it!</em> of 6 cognitive processes (excluded estimation) and 3 metacognitive processes. Participants received 8-10 sessions in total. Intervention consisted of Phase 1 (one-step problems) and Phase 2 (two-step problems). Participants entered Phase</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age: 15 – 17 years</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Grade: Grade 10 (n=2), Grade 9 (n=1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ethnicity: Hispanic (n=3)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Setting: schools in Florida, USA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Sample</td>
<td>Intervention</td>
<td>Measures</td>
<td>Main Findings</td>
</tr>
<tr>
<td>-------</td>
<td>--------</td>
<td>--------------</td>
<td>----------</td>
<td>---------------</td>
</tr>
<tr>
<td>Montague, Enders &amp; Dietz (2011)</td>
<td>2 groups of 20 schools each (after attrition, n = 24)</td>
<td>2 when they achieved 4 out of 5 problems correct in Phase 1.</td>
<td>and 4 weeks later, probe comprising 10 word problems (5 one- and 5 two-step problems)</td>
<td>Design: Randomised block design</td>
</tr>
<tr>
<td></td>
<td>Intervention (n = 319)</td>
<td></td>
<td></td>
<td>Intervention: Solve it! implemented across the academic year. 3 days of intensive instruction followed by weekly problem-solving practice sessions</td>
</tr>
<tr>
<td></td>
<td>Comparison (n = 460)</td>
<td></td>
<td></td>
<td>Comparison: teachers were asked to focus on problem solving for one class period per week</td>
</tr>
<tr>
<td></td>
<td>• 3 ability levels:</td>
<td>• Curriculum Based Measures (CBM): 30 math word problem sums selected from the Solve it! manual. Each measure consisted of 10 one-, two-, and three-step problems. There were 7 measures in total.</td>
<td>• Single FCAT Math problem taken from previous FCAT Math</td>
<td>o Learning disability (n = 78)</td>
</tr>
<tr>
<td></td>
<td>o Low achieving (n = 344)</td>
<td>o CBMs were administered monthly for intervention group</td>
<td></td>
<td>o Average achieving (n = 357)</td>
</tr>
<tr>
<td></td>
<td>o Gender: M = 46%, F = 54%</td>
<td>o For comparison group, measures were only administered at T1 (baseline, prior to intervention), T3, T5, T7 (post-intervention).</td>
<td></td>
<td>• Grade: 8</td>
</tr>
<tr>
<td></td>
<td>• Ethnicity: White 7%, Hispanic 64%, African American 27%</td>
<td></td>
<td></td>
<td>• Setting: middle schools in Florida, USA</td>
</tr>
<tr>
<td></td>
<td>• Free/reduced lunch: 72%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Setting: middle schools in Florida, USA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Sample</td>
<td>Intervention</td>
<td>Measures</td>
<td>Main Findings</td>
</tr>
<tr>
<td>-----------------------</td>
<td>------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
| Montague, Krawec, Enders & Dietz (2014) | • 2 groups of 20 schools each (after attrition, n = 34)  
  o Intervention (n = 644)  
  o Comparison (n = 415)  
• 3 ability levels:  
  o Learning disability (n = 86)  
  o Low achieving (n = 710)  
  o Average achieving (n = 263)  
• Gender: M = 43%, F = 57%  
• Grade: 7  
• Ethnicity: White 5%, Hispanic 65%, African American 30%  
• Free/reduced lunch: 80%  
• Setting: middle schools in Florida, USA | Design: Randomised block design  
Intervention: *Solve it!* implemented across the academic year. 3 days of intensive instruction followed by weekly problem-solving practice sessions  
Comparison: teachers were asked to focus on problem solving for one class period per week | • Curriculum Based Measures (CBM): same as in Montague et al. (2011) study  
• FCAT Grade 7 Math exam: assesses ratios and proportional relationships, geometry and measurement, numbers and operations, and base ten numbers  
• FCAT Grade 7 Reading exam: assesses vocabulary, reading, literary analysis, and the identification of specific information in reading passages | • There was a significant effect of the intervention on CBM, i.e. intervention group improved at a significantly higher rate than comparison group.  
• The growth rate for all three ability levels were similar.  
• There was insufficient evidence to support an intervention effect on FCAT Math and Reading. |
<p>| Whitby (2012) | 3 adolescents diagnosed with Autism Spectrum | Design: Multiple baseline across participants | • Curriculum Based Measures (CBM): 4 | • All three participants improved in their ability |</p>
<table>
<thead>
<tr>
<th>Study</th>
<th>Sample</th>
<th>Intervention</th>
<th>Measures</th>
<th>Main Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disorder</td>
<td>• Gender: Male (n=3)</td>
<td>Intervention: <em>Solve it!</em> implemented by a certified exceptional educator for 10 sessions – 5 sessions in training phase and 5 for acquisition phase. Participants progress to acquisition phase after they have acquired 100% mastery of strategy memorisation.</td>
<td>word problems taken from the <em>Solve it!</em> manual.</td>
<td>to solve math word problems. However, they did not maintain use of the strategy, i.e. there was overlap between data points at baseline and maintenance.</td>
</tr>
<tr>
<td></td>
<td>• Age: 13.7 – 14.3 years</td>
<td></td>
<td>• Each probe consisted of 5 word problems: CBM and 1 FCAT Math Grade 6 word problem. This was administered at each session.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Grade: Grade 7 (n=2) Grade 8 (n=1)</td>
<td></td>
<td>• Generalisation: Single FCAT Math Grade 6 problem of medium difficulty was also administered each week to participant and participants’ peers.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Ethnicity: not reported</td>
<td></td>
<td>• Follow-up: administered 4.5 weeks later, probe comprising 5 word problems administered daily for 3 days</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Setting: two public middle schools in Florida, USA.</td>
<td></td>
<td>• By acquisition phase, all three participants scored a higher percentage correct than their peer average. 2 participants also scored higher than their peer average during maintenance.</td>
<td></td>
</tr>
</tbody>
</table>

<p>| Zhu (2015)    | • 2 groups in 4 classes                       | Design: Randomised block design | 12 math word problems (none of which were used during instruction): represented four types of problems that were taught during instruction. The same measure was | Both intervention and comparison group performed better at post-test compared to pre-test. |
|               |   o intervention (n = 75)                     | Intervention: <em>Solve it!</em> implemented as an additional curricular complement to regular classes. Intervention lasted |                                                                              |                                                                              |
|               |   o comparison (n = 75)                       |              |                                                                              |                                                                              |
|               | • 4 ability levels:                           |              |                                                                              |                                                                              |
|               |   o Math Disability (MD) only (n = 16)        |              |                                                                              |                                                                              |
|               |   o Math and reading disability (MD/RD)       |              |                                                                              |                                                                              |
|               |                                                |              |                                                                              |                                                                              |</p>
<table>
<thead>
<tr>
<th>Study Sample</th>
<th>Intervention</th>
<th>Measures</th>
<th>Main Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n = 19)</td>
<td>for 8 weeks, with two 40-min sessions per week.</td>
<td>administered pre and post-intervention.</td>
<td>significantly greater than comparison group. This implies that the intervention has had a significant effect.</td>
</tr>
<tr>
<td>o Average achieving (n = 74)</td>
<td>Comparison: teachers were told to proceed with “business as usual” in those additional sessions.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>o High achieving (n = 41)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Gender: M = 58%, F = 42%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Ethnicity and free/reduced lunch: not reported</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Grade: 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Setting: elementary school in Henan, China</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• The growth rate also differed by ability levels. For the intervention group, students with MD only improved more than students with both MD/RD.</td>
</tr>
</tbody>
</table>
## Appendix B: Studies that were excluded after full text screening

<table>
<thead>
<tr>
<th>Excluded Studies</th>
<th>Rationale for exclusion</th>
</tr>
</thead>
</table>
Appendix C: Weight of Evidence

Weight of Evidence (WoE) A – Quality of Methodology

Four studies utilised a group experimental design, and were evaluated using Gersten et al. (2005). Two studies utilised a single case design (SCD) and were evaluated using Horner et al. (2005). The coding procedures were adopted from Jitendra et al. (2015), which scored the Quality Indicators (QI) for group and SCD studies on a 3-point scale: 3 (indicator met), 2 (indicator partially met), 1 (indicator not met). The coding procedures were deemed suitable as Jitendra et al.’s paper was a systematic review evaluating the evidence base of strategy instruction for math word problem solving. This review is similar except that a different intervention is being evaluated.

Appendix C1 and C2 present the coding procedures for group design studies and SCD studies respectively, that were the basis for evaluating the six studies in this review. Similar to Jitendra et al. (2015), the desirable QIs for group studies were evaluated on a dichotomous scale. The coding protocols for each study are in Appendices C3-C8.

After coding each study, an average score is obtained by averaging the individual scores for each QI. The criteria in Table 1 was used to incorporate both the essential and desirable QIs for group studies in the WoE A rating. The weighting criteria for SCD studies is presented in Table 2.

<table>
<thead>
<tr>
<th>Weighting</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>Average score of at least 2.5 on essential QIs AND at least 2 desirable QIs met</td>
</tr>
<tr>
<td>Medium</td>
<td>Average score of between 1.5 – 2.4 on essential QIs AND at least 1 desirable QI met</td>
</tr>
<tr>
<td>Low</td>
<td>Average score of less than 1.5, no desirable QI met</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Weighting</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>Average score of at least 2.5</td>
</tr>
<tr>
<td>Medium</td>
<td>Average score of between 1.5 – 2.4</td>
</tr>
<tr>
<td>Low</td>
<td>Average score of less than 1.5</td>
</tr>
</tbody>
</table>

The WoE A rating for each study is presented below.

<table>
<thead>
<tr>
<th>Study</th>
<th>Average score on Essential QIs</th>
<th>Number of Desirable QIs met</th>
<th>WoE A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Babakhani (2011)</td>
<td>1.5</td>
<td>0</td>
<td>Low</td>
</tr>
<tr>
<td>Coughlin &amp; Montague (2010)</td>
<td>2.4</td>
<td>NA</td>
<td>Medium</td>
</tr>
</tbody>
</table>
Weight of Evidence B – Relevance of Methodology

**Rationale**

WoE B considers whether the research design is suitable for answering the review question. As the purpose of the review is to evaluate the effectiveness of Solve it!, the WoE B criteria is based on Brannen’s (1992) evidence hierarchies. These hierarchies typically place randomised control trials and those with an active comparison group higher up, while case control and cohort studies rank lower in the hierarchy.

Horner et al. (2005) argued that single case designs are rigorous methodologies that can be used to establish evidence-based practice. However, the small sample sizes used in SCD studies mean that results ought to be generalised with caution. This is reflected in the WoE B criteria – SCD studies were able to receive up to a ‘medium’ rating.

**Table 3. WoE B Criteria**

<table>
<thead>
<tr>
<th>Weighting</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>For group design studies, all of the following criteria are met:</td>
</tr>
<tr>
<td>(3 points)</td>
<td>• Random assignment of participants to groups or randomised block design that demonstrates group equivalence</td>
</tr>
<tr>
<td></td>
<td>• Use of an active comparison group</td>
</tr>
<tr>
<td></td>
<td>• Measures used are reported to have high reliability (at least 0.70 for all measures used; Kratochwill, 2003)</td>
</tr>
<tr>
<td>Medium</td>
<td>• For group design studies, at least 2 of the criteria above are met.</td>
</tr>
<tr>
<td>(2 points)</td>
<td>• For SCD studies, multiple-baseline design across participants that provided at least three demonstrations of intervention effect.</td>
</tr>
<tr>
<td>Low</td>
<td>• For group design studies, at least 1 of the criteria above are met.</td>
</tr>
<tr>
<td>(1 point)</td>
<td>• For SCD studies, studies that did not provide three demonstrations of intervention effect.</td>
</tr>
</tbody>
</table>

The WoE B rating for each study is presented below.

<table>
<thead>
<tr>
<th>Study</th>
<th>Randomisation</th>
<th>Comparison group</th>
<th>Measures</th>
<th>WoE B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Babakhani (2011)</td>
<td>Unclear</td>
<td>Not reported</td>
<td>Y</td>
<td>Low</td>
</tr>
</tbody>
</table>
Weight of Evidence C – Relevance of Evidence to Review Question

Rationale

WoE C considers the relevance of the focus of the evidence to the review question, which is to evaluate whether *Solve it!* improved the mathematics attainment of students with math difficulties. As the population of focus is students with math difficulties, an important criteria was whether the study sample included students with demonstrated math difficulties. To determine the effectiveness of *Solve it!* it would also be important to ensure that the intervention is implemented as intended. To evaluate *Solve it!*’s effectiveness as an educational intervention, the use of multi-source and/or multi-method outcome measures that demonstrate a generalised and sustained effect on math attainment are also desired.

Table 4. WoE C Criteria

<table>
<thead>
<tr>
<th>Weighting</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>High</strong> (3 points)</td>
<td>To receive a ‘high’ weighting, the study must fulfil all of the following criteria:</td>
</tr>
<tr>
<td></td>
<td>• Participants: Study provides evidence to show that participants have math difficulties through standardised assessments (Gersten et al., 2005)</td>
</tr>
<tr>
<td></td>
<td>• Intervention: Study implements intervention in its entirety, as described in the introduction, i.e. seven cognitive and three metacognitive processes</td>
</tr>
<tr>
<td></td>
<td>• Generalisation: Use of multiple outcome measures, including a general measure of math attainment</td>
</tr>
<tr>
<td></td>
<td>• Maintenance: Follow-up measure administered more than 2 weeks after intervention (Gersten et al., 2005)</td>
</tr>
<tr>
<td><strong>Medium</strong> (2 points)</td>
<td>To receive a ‘medium’ weighting, the study must fulfil at least 2 of the criteria listed above.</td>
</tr>
<tr>
<td><strong>Low</strong> (1 point)</td>
<td>To receive a ‘low’ weighting, the study fulfils at least 1 of the criteria listed above.</td>
</tr>
</tbody>
</table>
The WoE C rating for each study is presented below.

<table>
<thead>
<tr>
<th>Study</th>
<th>Participants</th>
<th>Intervention</th>
<th>Generalisation</th>
<th>Maintenance</th>
<th>WoE C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Babakhani (2011)</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td>Low</td>
</tr>
<tr>
<td>Coughlin &amp; Montague (2010)</td>
<td>Y</td>
<td></td>
<td>Y</td>
<td></td>
<td>Med</td>
</tr>
<tr>
<td>Montague et al. (2011)</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
<td>Med</td>
</tr>
<tr>
<td>Montague et al. (2014)</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
<td>Med</td>
</tr>
<tr>
<td>Whitby (2012)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td>Med</td>
</tr>
<tr>
<td>Zhu (2015)</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td>Low</td>
</tr>
</tbody>
</table>

**Weight of Evidence D — Overall Weight of Evidence**

The WoE D rating of each study is obtained by averaging the WoE ratings for A, B and C. Table 5 describes the weighting criteria and Table 6 summarises the WoE ratings for each study.

*Table 5. Weighting criteria for WoE D*

<table>
<thead>
<tr>
<th>Weighting</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>Average score of at least 2.5</td>
</tr>
<tr>
<td>Medium</td>
<td>Average score of between 1.5 and 2.4</td>
</tr>
<tr>
<td>Low</td>
<td>Average score of less than 1.4</td>
</tr>
</tbody>
</table>

*Table 6. Summary of WoE A – D for each study in this review*

<table>
<thead>
<tr>
<th>Study</th>
<th>Quality of methodology (A)</th>
<th>Relevance of methodology (B)</th>
<th>Relevance of evidence to the review question (C)</th>
<th>Overall weight of evidence (D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Babakhani (2011)</td>
<td>1.5</td>
<td>1.0</td>
<td>1.0</td>
<td>Low (1.2)</td>
</tr>
<tr>
<td>Coughlin &amp; Montague (2010)</td>
<td>2.4</td>
<td>2.0</td>
<td>2.0</td>
<td>Med (2.1)</td>
</tr>
<tr>
<td>Montague et al. (2011)</td>
<td>2.3</td>
<td>1.0</td>
<td>2.0</td>
<td>Med (1.8)</td>
</tr>
<tr>
<td>Montague et al. (2014)</td>
<td>2.8</td>
<td>3.0</td>
<td>2.0</td>
<td>High (2.6)</td>
</tr>
<tr>
<td>Study</td>
<td>Quality of methodology (A)</td>
<td>Relevance of methodology (B)</td>
<td>Relevance of evidence to the review question (C)</td>
<td>Overall weight of evidence (D)</td>
</tr>
<tr>
<td>------------------</td>
<td>-----------------------------</td>
<td>-----------------------------</td>
<td>--------------------------------------------------</td>
<td>--------------------------------</td>
</tr>
<tr>
<td>al. (2014)</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>Med (2.2)</td>
</tr>
<tr>
<td>Whitby (2012)</td>
<td>2.7</td>
<td>2.0</td>
<td>2.0</td>
<td>Med (2.2)</td>
</tr>
<tr>
<td>Zhu (2015)</td>
<td>2.4</td>
<td>3.0</td>
<td>1.0</td>
<td>Med (2.1)</td>
</tr>
</tbody>
</table>

**References**

# Appendix C1: Coding Procedures for Essential Quality Indicators of Group Design Studies

<table>
<thead>
<tr>
<th>Essential Quality Indicator</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description of Participants</td>
<td>Somewhat</td>
</tr>
<tr>
<td>Information on participants’ disability or difficulties (e.g., age, race, gender, IQ, socioeconomic status, English language learner, scores on academic assessments)</td>
<td>Cited school district/state criteria for disability status; did not document specific difficulties using assessments or diagnostic criteria</td>
</tr>
</tbody>
</table>

<p>| Equivalence of groups across conditions | Did not randomly assign participants or classrooms to conditions AND did not document comparability of participants in conditions on a mathematics measure (did not provide the necessary scores for the reader to be able to assess equivalence) | Randomly OR nonrandomly assigned participants or classrooms to conditions AND documented comparability of participants in conditions on at least two demographic variables, as well as a mathematics measure (or provided the necessary scores for the reader to be able to assess equivalence) | Randomly assigned participants or classrooms to conditions AND documented comparability of participants in conditions on at least three demographic variables, as well as a mathematics measure (or provided the necessary scores for the reader to be able to assess equivalence) |</p>
<table>
<thead>
<tr>
<th>Essential Quality Indicator</th>
<th>No</th>
<th>Somewhat</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>necessary scores for the reader to be able to assess equivalence</td>
<td>for the reader to be able to assess equivalence</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Information on intervention agents (e.g., years of experience, teaching certificates, level of education, age, gender, race, and familiarity with the intervention); equivalence of intervention agents across conditions

Specified intervention agents for each condition, but did not provide descriptive information OR did not specify intervention agents for each condition

Intervention agent was same for all conditions OR specified intervention agents for each condition and provided some descriptive information

Described intervention agents and randomly assigned or counterbalanced them across conditions OR documented comparability of intervention agents in conditions on at least three relevant characteristics

---

**Description and implementation of intervention and comparison conditions**

**Description of intervention** (e.g., conceptual underpinnings, duration of intervention, detailed instructional procedures, teacher actions and language, use of instructional materials, and student behaviors)

Provided specific information on two or fewer relevant dimensions of the intervention

Provided specific information on at least three relevant dimensions of the intervention, OR directed readers to another article for description of procedures

Provided specific information on at least four relevant dimensions of the intervention

**Description and measurement of procedural fidelity**

Provided no description of treatment fidelity

Provided description of treatment fidelity (e.g., instruction provided by using scripted lessons)

Described treatment fidelity and assessed the extent to which specific components of the intervention were implemented (e.g., checklists of intervention components completed by an...
<table>
<thead>
<tr>
<th>Essential Quality Indicator</th>
<th>No 1</th>
<th>Somewhat 2</th>
<th>Yes 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>observer, self-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>monitoring checklists, or</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>analysis of videotapes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>and field notes)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Did not describe nature</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>of instruction in</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>comparison conditions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Described instruction on</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>at least two relevant</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dimensions (e.g.,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>use of instructional</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>materials, grouping,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>setting, and time for</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>instruction)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Did not describe nature</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>of instruction in</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>comparison conditions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Described nature of</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>instruction, specifically</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>teacher action and</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>expected student</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>behaviours</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Outcome measures</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple measures</td>
<td>Employed only outcome measures aligned with the intervention</td>
<td>Employed only measures of generalised performance</td>
<td>Employed outcome measures aligned with the intervention AND measures of generalised performance</td>
</tr>
<tr>
<td>or measures of</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>generalised</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>performance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Measured more</td>
<td>Measured within 1 month of intervention</td>
<td>Measured within 2 weeks of intervention</td>
<td></td>
</tr>
<tr>
<td>than 1 month of</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intervention</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Appropriateness of | | | |
| time of data       | | | |
| collection         | | | |
| Measured more than | Measured within 1 month of intervention | Measured within 2 weeks of intervention |
| 1 month of          | | | |
| intervention        | | | |

<table>
<thead>
<tr>
<th>Data Analysis</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Did not align data analysis techniques with the research questions/hypotheses and did not use appropriate unit of analysis</td>
<td>Aligned data analysis techniques with the research questions/hypotheses, but did not use appropriate unit of analysis</td>
<td>Aligned data analysis techniques with the research question/hypotheses and used appropriate unit of analysis</td>
<td></td>
</tr>
<tr>
<td>Effect sizes</td>
<td>Effect size not reported in text</td>
<td>Effect size reported in text but not interpreted</td>
<td>Effect size reported in text and interpreted</td>
</tr>
</tbody>
</table>
### Appendix C2: Coding Procedures for Single Case Design Studies

<table>
<thead>
<tr>
<th></th>
<th>No (1)</th>
<th>Somewhat (2)</th>
<th>Yes (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Participants and Setting</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant description (e.g., age, gender, IQ, disability, diagnosis)</td>
<td>Did not provide operational definition or criteria for disability; fewer than three details of participants included</td>
<td>Provided operational definition or criteria for disability; provided information on three demographic variables</td>
<td>Provided operational definition of disability; provided information on four demographic variables</td>
</tr>
<tr>
<td>Participant selection</td>
<td>Not described; included mathematics preassessment data OR described a criterion for selection of participants; mathematics preassessment data are not included</td>
<td>Described a criterion for selection of participants; included mathematics preassessment data</td>
<td>Described precise criteria (e.g., deficient mathematics performance) for selection of participants; included mathematics preassessment data</td>
</tr>
<tr>
<td>Setting description (e.g., type of classroom, room arrangement, number of students to teachers)</td>
<td>Not described OR described a few critical features of setting</td>
<td>Described some critical features of setting</td>
<td>Precisely described critical features of setting to allow replication</td>
</tr>
<tr>
<td><strong>Dependent Variable</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Description of DV</td>
<td>Described subjectively or globally OR is not described</td>
<td>Described adequately, but not in operational terms</td>
<td>Described with operational precision to allow direct observation and replication</td>
</tr>
<tr>
<td>Measurement procedure</td>
<td>Measurement procedure does not generate a quantifiable index</td>
<td>Measurement procedure generates a quantifiable index for some but not all variables of</td>
<td>Measurement procedure generates a quantifiable index for all variables of</td>
</tr>
<tr>
<td>No 1</td>
<td>Somewhat 2</td>
<td>Yes 3</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>------------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td><strong>variables of interest</strong></td>
<td>interest</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Measurement validity and description</strong></td>
<td>Measurement is not valid; minimal or no description of procedure</td>
<td>Measurement is valid; limited description of the procedure</td>
<td>Measurement is valid; precise description of procedure to allow replication</td>
</tr>
<tr>
<td><strong>Measurement frequency</strong></td>
<td>Measurement is infrequent (fewer than three data points per condition)</td>
<td>Measurement is repeated frequently (three data points for most conditions) <strong>OR</strong> criterion performance is met and documented with two data points <strong>AND</strong> overall pattern of performance for most but not all conditions is established</td>
<td>Measurement is repeated frequently, with a minimum of three data points per condition <strong>OR</strong> criterion performance is met and documented with two data points <strong>AND</strong> overall pattern of performance for most but not all conditions is established</td>
</tr>
<tr>
<td><strong>Measurement reliability</strong></td>
<td>Reliability data (IOA) are provided for some, but not all DVs (or reported as only one score across all measures), but does not meet minimum standards <strong>OR</strong> reliability data (IOA) are not provided for any of the DVs</td>
<td>Reliability data (IOA) are provided for some, but not all DVs (or reported as only one score across all measures); meets minimum standards <strong>OR</strong> reliability data (IOA) are provided for each DV across participants or behaviors in each condition/phase, but does not meet minimum standards</td>
<td>Reliability data (IOA) are provided for each DV across participants or behaviors in each condition/phase; meets minimum standards (IOA=80%)</td>
</tr>
</tbody>
</table>

**Independent Variable**

<p>| Description of IV (e.g., instructional) | Provided specific information on two | Provided specific information on at | Provided specific information on at |</p>
<table>
<thead>
<tr>
<th></th>
<th>No 1</th>
<th>Somewhat 2</th>
<th>Yes 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>materials, procedures, length of session, duration of intervention)</td>
<td>or fewer relevant dimensions of the intervention</td>
<td>least three relevant dimensions of the intervention</td>
<td>least four relevant dimensions of the intervention</td>
</tr>
<tr>
<td>Manipulation of IV</td>
<td>IV is manipulated, but there is no documentation of experimental control</td>
<td>IV is manipulated, but documentation of experimental control is not precise (i.e., the researcher does not document when and how the IV conditions change)</td>
<td>IV is systematically manipulated with precise documentation of experimental control (i.e., researcher documents when and how the IV conditions change)</td>
</tr>
<tr>
<td>Fidelity of implementation</td>
<td>Procedural fidelity is not reported</td>
<td>Procedural fidelity is reported (use of teaching scripts), but not directly measured</td>
<td>Procedural fidelity is reported by direct measurement of IV</td>
</tr>
<tr>
<td>Baseline Measurement of DV</td>
<td>DV is measured infrequently (based on fewer than three data points) in baseline, OR DV is measured frequently (based on a minimum of three data points), but baseline is not stable and/or in the expected direction prior to intervention implementation for the majority of participants or behaviors</td>
<td>DV is measured frequently (based on a minimum of three data points), AND baseline is stable and/or in the expected direction prior to intervention implementation for the majority but not all participants or behaviors</td>
<td>DV is measured frequently (based on a minimum of three data points), AND baseline is stable and/or in the expected direction prior to intervention implementation for all participants or behaviors</td>
</tr>
<tr>
<td>Description of baseline condition (e.g., materials, procedures,</td>
<td>Description of baseline condition is imprecise, general, or is not</td>
<td>Description of baseline condition is adequate, but is missing some</td>
<td>Description of baseline condition is precise to allow replication</td>
</tr>
</tbody>
</table>

42
<table>
<thead>
<tr>
<th>No</th>
<th>Somewhat</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

**Experimental control/internal validity**

<table>
<thead>
<tr>
<th>Setting</th>
<th>Provided details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental effect (when predicted change in the DV covaries with manipulation of the IV [i.e., change in response patterns when the problem solving intervention is implemented])</td>
<td>The data across all phases of the study document one or no demonstration of an experimental effect</td>
</tr>
</tbody>
</table>

**Internal validity**

<table>
<thead>
<tr>
<th>Design controls for few threats to internal validity</th>
<th>Design controls for some threats to internal validity</th>
<th>Design controls for most threats to internal validity</th>
</tr>
</thead>
</table>

**Results** (interpretation of data in terms of level [more than small mean changes across phases], trend, variability of performance as well as immediacy of effect, minimal overlap of data points across adjacent phases, and consistency of data patterns across similar phases)

<p>| Insufficient data points AND/OR pattern of results does not demonstrate experimental control | Sufficient data points within each phase present AND pattern of results demonstrate some experimental control in terms of meeting at least four features for all or the majority of participants or behaviors | Sufficient data points within each phase present AND pattern of results demonstrate experimental control in terms of meeting the following features: (a) level, (b) trend, (c) variability, (d) immediacy of effects evident following the onset of the intervention, (e) minimal overlap of data points in |</p>
<table>
<thead>
<tr>
<th><strong>No</strong></th>
<th><strong>Somewhat</strong></th>
<th><strong>Yes</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>document experimental control – data pattern in one phase [intervention] differs more than would be expected from the data pattern observed or extrapolated from the previous phase [baseline])</td>
<td>adjacent phases, and (f) consistent data patterns across similar phases for all participants or behaviors</td>
<td></td>
</tr>
</tbody>
</table>

### External validity

<table>
<thead>
<tr>
<th>Action</th>
<th>No replications</th>
<th>Few replications across different participants, different conditions, and/or different measures of the DV</th>
<th>Three or more replications across different participants, different conditions, and/or different measures of the DV</th>
</tr>
</thead>
</table>

### Social validity

<table>
<thead>
<tr>
<th>Importance</th>
<th>Not important</th>
<th>Somewhat important</th>
<th>Important</th>
</tr>
</thead>
<tbody>
<tr>
<td>DV</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Change in DV (e.g., mean level, PND)</td>
<td>Not socially important</td>
<td>Somewhat socially important</td>
<td>Socially important</td>
</tr>
</tbody>
</table>

<p>| Implementation of IV is practical and cost-effective | Social validity data about intervention procedures are not gathered from intervention agents or students | Social validity data provide documentation of 1-2 features (acceptability, feasibility, effectiveness, and continued use) | Social validity data provide documentation of at least three features (acceptability, feasibility, effectiveness, and continued use) |
|-----------------------------------------------------|--------------------------------------------------|--------------------------------------------------|
| Nature of implementation of IV | Not reported or only one feature (i.e., typical) | At least two features (i.e., typical intervention) | IV implemented by (a) typical intervention |</p>
<table>
<thead>
<tr>
<th></th>
<th>No 1</th>
<th>Somewhat 2</th>
<th>Yes 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>intervention agents, typical settings, or over an extended time period) of IV implementation is documented</td>
<td>agents, typical settings, or over an extended time period) of IV implementation are documented</td>
<td>agents, (b) in typical settings, (c) for an extended time period</td>
<td></td>
</tr>
</tbody>
</table>
Appendix C3: Coding Protocol for Babakhani (2011)

Date: 30/01/2016


Intervention Name (description of study): Teaching cognitive and meta-cognitive strategies to Grade 4 students with verbal math problem solving difficulties

Research design: Randomised block design

Type of Publication: Journal Article

Gersten et al., (2005). Quality Indicators for Group Experimental and Quasi-Experimental Research in Special Education

Essential Quality Indicators

Describing Participants

Was sufficient information provided to determine/confirm whether the participants demonstrated the disability(ies) or difficulties presented? ☐ Yes ☒ Somewhat Although the author defined verbal math problem solving difficulties as between average and 1 SD below average on author’s VMPS test; there was little information provided about this test. Also did not provide information on 4 demographic variables. ☐ No ☐ Unknown/Unable to Code

Were appropriate procedures used to increase the likelihood that relevant characteristics of participants in the sample were comparable across conditions? ☐ Yes ☒ Somewhat It was reported that participants were randomly sampled, matched on intelligence, age and math performance, but no procedure or evidence was documented ☐ No ☐ Unknown/Unable to Code

Was sufficient information given characterizing the interventionists or teachers provided? Did it indicate whether they were comparable across conditions? ☐ Yes ☐ Somewhat ☒ No Interventionist was not specified for control condition, and background information on interventionist for treatment condition was not described
Mean = 1.67

Implementation of the Intervention and Description of Comparison Conditions

Was the intervention clearly described and specified?
☐ Yes
☒ Somewhat Conceptual underpinnings, teacher actions and language, and student behaviours described. However, it was insufficient to obtain a clear picture of how the instruction was conducted e.g. duration of intervention was not specified.
☐ No
☐ Unknown/Unable to Code

Was the fidelity of implementation described and assessed?
☐ Yes
☐ Somewhat
☒ No
☐ Unknown/Unable to Code

Was the nature of services provided in comparison conditions described?
☐ Yes
☐ Somewhat
☒ No
☐ Unknown/Unable to Code

Mean = 1.33

Outcome Measures

Were multiple measures used to provide an appropriate balance between measures closely aligned with the intervention and measures of generalised performance?
☐ Yes
☐ Somewhat
☒ No Only one measure of 10 verbal math word problems used as DV
☐ Unknown/Unable to Code

Were outcomes for capturing the intervention’s effect measured at the appropriate times?
☒ Yes The instructor administered a post-test two weeks after intervention.
☐ No
☐ N/A
☐ Unknown/Unable to Code

Mean = 2.00

Data Analysis
Were the data analysis techniques appropriately linked to key research questions and hypotheses? Were they appropriately linked to the unit of analysis in the study?

☐ Yes
☐ Somewhat
☒ No The data analysis technique did not take into account other possible co-variants that could account for difference in scores e.g., age, intelligence, math performance. Also t-test results are not reported.
☐ Unknown/Unable to Code

Did the research report include not only inferential statistics but also effect size calculations?

☐ Yes
☐ Somewhat
☒ No
☐ Unknown/Unable to Code

**Mean = 1.00**

**Desirable Quality Indicators**

Was data available on attrition rates among intervention samples? Was severe overall attrition documented? If so, is attrition comparable across samples? Is overall attrition less than 30%?

☐ Yes
☐ No
☐ N/A
☒ Unknown/Unable to Code

Did the study provide not only internal consistency reliability but also test-retest reliability and interrater reliability (when appropriate) for outcome measures? Were data collectors and/or scorers blind to study conditions and equally (un)familiar to examinees across study conditions?

☐ Yes
☒ No
☐ N/A
☐ Unknown/Unable to Code

Were outcomes for capturing the intervention’s effect measured beyond an immediate posttest?

☐ Yes
☒ No
☐ N/A
☐ Unknown/Unable to Code

Was evidence of the criterion-related validity and construct validity of the measures provided?
Did the research team assess not only surface features of fidelity implementation (e.g. number of minutes allocated to the intervention or teacher/interventionist following procedures specified), but also examine quality of implementation?
☐ Yes
☒ No
☐ N/A
☐ Unknown/Unable to Code

Was any documentation of the nature of instruction or series provided in comparison conditions?
☐ Yes
☒ No
☐ N/A
☐ Unknown/Unable to Code

Did the research report include actual audio or videotape excerpts that capture the nature of the intervention?
☐ Yes
☒ No
☐ N/A
☐ Unknown/Unable to Code

Were results presented in a clear, coherent fashion?
☐ Yes
☒ No
☐ N/A
☐ Unknown/Unable to Code

Average WoE A across the 4 judgement areas:
Sum of X / N = (1.67+1.33+2+1)/4 = 1.5
X = average score for each judgement area
N = number of judgement areas

Date: 30/01/2016


Intervention Name (description of study): Solve It! cognitive strategy instruction for three adolescents with Spina Bifida

Research design: Multiple-baseline across participants

Type of Publication: Journal Article

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

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 Demographic information provided in Table 1
☐ Somewhat
☐ No
☐ Unknown/Unable to Code

The process for selecting participants is described with replicable precision.
☒ Yes Inclusion criteria is described clearly in Participants section
☐ Somewhat
☐ No
☐ Unknown/Unable to Code

Critical features of the physical setting are described with sufficient precision to allow replication.
☐ Yes
☒ Somewhat Setting for each participant was described in Procedures section – classroom described, and later on also reported that researcher was instructor, but assumed that it is individually conducted?
☐ No
☐ Unknown/Unable to Code

Mean = 2.67

Dependent Variable

Dependent variables are described with operational precision.
☒ Yes DV is percent correct on word problems
Each dependent variable is measured with a procedure that generates a quantifiable index.
☒ Yes Measurement of DV is percent correct on word problems.
☐ No
☐ N/A
☐ Unknown/Unable to Code

Measurement of the dependent variable is valid and described with replicable precision.
☒ Yes It was mentioned that these problems had been previously validated as part of the Solve it! intervention. Procedure of administration is clear.
☐ No
☐ N/A
☐ Unknown/Unable to Code

Dependent variables are measured repeatedly over time.
☒ Yes Mathematical word problem probes were administered at each session and phase. Minimum of 3 data points at each phase.
☐ No
☐ N/A
☐ Unknown/Unable to Code

Data are collected on the reliability or inter-observer agreement associated with each dependent variable, and IOA levels meet minimal standards (e.g. IOA = 80%; Kappa = 60%)
☐ Yes
☐ Somewhat
☒ No Data was not provided on reliability or inter-observer agreement of the DV.
☐ Unknown/Unable to Code

Mean = 2.60

Independent Variable

Independent variable is described with replicable precision.
☒ Yes It is reported that materials from the Solve It! curriculum were used, which included scripted lessons, etc. Number of sessions and how many problems they complete in each session is clear. However, length of session and duration of intervention was unclear e.g. daily?
☐ Somewhat
☐ No
☐ Unknown/Unable to Code
Independent variable is systematically manipulated and under the control of the experimenter.
☒ Yes Documentation in Procedures section
☐ No
☐ N/A
☐ Unknown/Unable to Code

Overt measurement of the fidelity of implementation for the independent variable is highly desirable.
☐ Yes
☒ Somewhat Use of scripted lessons and researchers concurred on the basis of notes taken that it was implemented with 100% fidelity.
☐ No
☐ Unknown/Unable to Code

Mean = 2.67

Baseline

The majority of single-subject research studies will include a baseline phase that provides repeated measurement of a dependent variable and establishes a pattern of responding that can be used to predict the pattern of future performance, if introduction or manipulation of the independent variable did not occur.
☐ Yes
☒ Somewhat Three baseline tests were administered to students over 1 month to establish baseline, but baseline for one participant seemed to be increasing
☐ N/A
☐ Unknown/Unable to Code

Baseline conditions are described with replicable precision.
☐ Yes
☐ Somewhat
☒ No Article does not describe baseline condition
☐ Unknown/Unable to Code

Mean = 1.50

Experimental Control/internal Validity

The design provides at least three demonstrations of experimental effect at three different points in time.
☒ Yes
☐ Somewhat
☐ No
☐ Unknown/Unable to Code
The design controls for common threats to internal validity (e.g., permits elimination of rival hypotheses).

☐ Yes
☒ Somewhat Use of multiple baseline design, but no measure of comparison with peer group
☐ No
☐ Unknown/Unable to Code

The results document a pattern that demonstrates experimental control.

☐ Yes
☒ Somewhat not immediate for one, some overlap, and some variability but overall trend and levels are consistent with experimental control
☐ No
☐ Unknown/Unable to Code

**Mean = 2.33**

**External Validity**

Experimental effects are replicated across participants, settings, or materials to establish external validity.

☒ Yes Effects were replicated across 3 adolescents with SBM
☐ Somewhat
☐ No
☐ Unknown/Unable to Code

**Mean = 3.00**

**Social Validity**

The dependent variable is socially important.

☒ Yes word problem sums are part of the curriculum, one participant also reported that the intervention helped improve his self-esteem
☐ No
☐ N/A
☐ Unknown/Unable to Code

The magnitude of change in the dependent variable resulting from the intervention is socially important.

☒ Yes
☐ No
☐ N/A
☐ Unknown/Unable to Code

Implementation of the independent variable is practical and cost effective.

☐ Yes
☒ Somewhat Some evidence from the participants that they benefited from the intervention
☐ No
☐ Unknown/Unable to Code

Social validity is enhanced by implementation of the independent variable over extended time periods, by typical intervention agents, in typical physical and social contexts.
☐ Yes
☐ No
☒ No Intervention occurred daily (possibly?) by the researcher (not known to the participants), in a separate classroom
☐ Unknown/Unable to Code

Mean = 2.25

Average WoE A across the 7 judgement areas:
Sum of X / N = (2.67+2.6+2.67+1.5+2.3+3+2.25)/7 = 2.4
X = average score for each judgement area
N = number of judgement areas