Cathryn Neesam

Case Study 1: An Evidence-Based Practice Review Report

Theme: School Based Interventions for Learning

Are concrete manipulatives effective in improving the mathematics skills of children with mathematics difficulties?

Summary

Concrete manipulatives can be defined as 'objects which can be touched and moved by students to introduce or reinforce a mathematical concept' (Hartshorn & Boren, 1990, p.2). For children having difficulty with maths, manipulatives can represent abstract concepts in a visual way that fosters understanding and provides a strategy from which to work. A recent meta-analysis (Carbonneau, Marley & Selig, 2013) found small to medium effects in favour of concrete manipulative use, but as yet, whether they are an effective strategy for children with maths difficulties has not been reviewed. This review aims to establish whether there is an evidence base to support the use of manipulatives in improving the mathematics skills of children with mathematics difficulties. Findings showed promising evidence in favour of manipulatives for children with maths difficulties, providing they feature certain key components. Limitations and future research directions are discussed.

Introduction

Manipulative materials have been defined as "objects which can be touched and moved by students to introduce or reinforce a mathematical concept" (Hartshorn & Boren, 1990, p.2). They provide a concrete learning experience whereby physical objects are manipulated in order to develop meaningful comprehension of the symbols they represent (Underhill, 1981). Examples range from everyday objects such as buttons and coins to objects specifically designed for use in maths such as Numicon, base-ten blocks and Unifix cubes (Domino, 2010). At a basic level, beads can be sorted into groups to support understanding of basic arithmetic functions like multiplication and division. With more complex functions for example, plastic algebra tiles may be moved around to help solve equations. Virtual manipulatives, a computer generated form of the traditional resource, are also gaining popularity (e.g. Bouck, Satsangi, Doughty & Courtney, 2014), however this review will focus on the physical form due to their low cost and easy access in classrooms.

Whatever the object, ‘good manipulatives’ can be considered as those that “aid students in building, strengthening and connecting various representations of mathematical ideas” (Clements, 1999, p.50). The underlying practice - hands-on experience of manipulating objects – is simple, but has been applied to a range of maths concepts with promising results (Bouck et al., 2014) including place value (Peterson, Mercer & O’Shea, 1988), addition and subtraction (Fuson & Briars, 1990), word problem solving (Stellingwerf & Van Lieshout, 1999), fractions (Jordan, Miller & Mercer, 1999) and algebra (Maccini & Hughes, 2000).

Manipulatives have been used with whole classes (Witzel, 2005), in small groups and individually (Cass, Cates & Smith, 2003). They have also been used with students of a range of ages; in the US, manipulatives are actively recommended for use with all grade levels by the National Council of Teachers of Mathematics (NCTM, 2000). A popular approach emerging from this endorsement is the concrete-representational-abstract teaching sequence. This process provides explicit instruction with manipulatives then fades these aids until students solve problems fluently using numbers only (Flores, Hinton & Schweck, 2014).

Theoretical support for the use of manipulatives dates back to developmental theorists such as Piaget (1965) and Bruner (1964) who argued that children do not come into the world with the capacity for abstract thought, but must instead construct abstract concepts through interactions with objects in their environment. Manipulatives provide such an opportunity and bridge the gap between concrete and abstract (Domino, 2010). This theory would imply that younger children benefit most from the use of manipulatives, but as older children can also improve their maths skills using manipulatives, there must be additional factors to consider.

Others have suggested that manipulatives help children's maths development because they provide an additional resource; with multiple resources children are more likely to perform at optimal levels (Sternberg & Grigorenko, 2004) and the teacher increases the chances of reaching the wide range of students in their class. They enable children to draw on their real-world practical knowledge (Rittle-Johnson & Koedinger, 2005) and induce physical action, which has been shown to enhance memory and understanding (Martin & Schwartz, 2005).

However, the efficacy of manipulatives has been questioned, with some studies reporting detrimental effects (e.g. Amaya, Uttal, O'Doherty, Liu & DeLoache, 2007). McNeil and Jarvin (2007) have argued that the theory behind them may be counter-intuitive, as manipulatives require dual representation - as an object in its own right and as a symbol of a mathematical concept or procedure - which is cognitively demanding. Moyer (2001) found that teachers often used manipulatives to add variety or fun to lessons. If used ineffectively, they may have little impact on students' understanding (Thompson, 1992). Additionally, simply incorporating manipulatives into maths teaching may not be enough to increase achievement (Carbonneau et al., 2013); it cannot be assumed that children will immediately see the mathematical concepts or relationships by interacting with objects. Therefore, it is important that manipulatives are not used as an ‘add on’, but their use is explicitly explained and modelled to ensure understanding.

Compared to research on language difficulties, our understanding of maths difficulties is in its infancy (Chinn, 2014). Children with such difficulties may struggle to acquire number concepts, exhibit confusion over maths symbols, lack a grasp of numbers and have problems learning and remembering number facts (Reeve & Gray, 2013). Such difficulties can be pervasive as children progress through school and are required to tackle increasingly complex concepts such as fractions, probability and algebra (Bryant, Bryant, Shin & Pfannenstiel, 2013). Acquiring academic maths skills like addition and multiplication at an early age provide the academic foundation for later success in maths-based courses (Fletcher, Boon & Cihak, 2010). However there is a worrying issue in the UK of children leaving school having underachieved in maths to the point of being functionally innumerate, which is associated with NEET (not in education, employment or training) status (Rashid & Brooks, 2010). National Numeracy (2014) report that maths skills have worsened over time, with only 22% of adults in England having maths skills equivalent to grade C or above. Evidently, something needs to change about the way maths is taught to those with difficulties.

Unlike in the US, “practical resources [for example, concrete objects and measuring tools]” are only referred to in UK government guidance for use with children in Key Stage (KS) 1 and 2 of the National Curriculum (DfE, 2013, p.5). There is no mention of their benefits in KS3 or 4 guidance, yet research has found a positive impact on older children in more complex skills like algebra (Maccini & Hughes, 2000). Once children leave primary school they are expected to be competent in a range of more abstract maths concepts that would arguably be clearer if made concrete. Not only can manipulatives assist students in developing a deeper understanding of mathematical concepts, but they are a valuable means of engaging students in the language and communication of mathematical ideas (Domino, 2010). If manipulatives are found to be effective in improving the skills of children with maths difficulties, putting such interventions into place is likely to help children achieve at higher levels and leave school with the functional maths skills required in society.

Significantly more work has been completed over the years with children who are low-attaining in literacy than in maths, meaning when Educational Psychologists (EPs) become involved with children with maths difficulties, there is a smaller pool of research upon which to draw (Holmes & Dowker, 2013). To support schools to narrow the gap in maths attainment and reduce the numbers of children leaving school without functional maths skills, it is crucial that EPs recommend strategies and interventions that are evidence-based. A recent meta-analysis (Carbonneau et al., 2013) found small to medium effects in favour of the use of manipulatives compared to instruction that used only abstract maths symbols for school age children. However, the effectiveness specifically relating to their use with children with maths difficulties has not been reviewed. Therefore, this systematic review seeks to answer the following question:

Are concrete manipulatives effective in improving the mathematics skills of children with mathematics difficulties?

Critical review of the evidence base

Literature search

A comprehensive literature search was conducted on 17th December 2015 using the online databases ERIC (EBSCO), PsycINFO, PsychSource and Web of Science. Table 1 below shows the search terms used to identify articles relevant to the review question. A multi-field search was conducted using the commands 'OR' and 'AND' to combine search terms. The search was limited to 'title' and 'abstract' in all databases and only peer-reviewed articles were selected.

Table 1: search terms used to locate studies

Search terms applied				
math* OR numer*	AND	manipulative* OR concrete	AND	difficult* OR disabilit* OR weakness OR dyscalcul* OR special educational need*

* denotes truncated term, e.g. math* will search for mathematics, mathematical, maths etc.

As Figure 1 below shows, the initial search yielded 495 studies. Using the exclusion criteria detailed in Table 2, these were scrutinised to leave 5 studies included in this systematic review. The list of studies excluded at full text review with rationale can be found in Appendix 1.

Table 2: Inclusion and exclusion criteria

Criterion	Inclusion criteria	Exclusion criteria	Rationale
1) Publication type	a) Must be published in a peer reviewed journal	Study is not published in a peer reviewed journal	Peer review is an external and rigorous validation; the study should be of a high methodological quality
	b) Must be written in English	Study is not written in English	Resources for translation not available
2) Study type	a) Must report primary empirical data	Study does not report primary empirical data, e.g. reviews	Empirical data required for analysis to explore the impact of the intervention
	b) Must contain quantitative data	Study does not contain quantitative data, e.g. qualitative data	Quantitative data required to calculate effect sizes
	c) Must report child outcome measures	Does not report on child outcomes, e.g. teacher's evaluation	Required to answer review question
3) Participants	a) Participants must be school age	Participants do not attend school, e.g. adults or university students	Focus of the review is on the use of manipulatives in school settings
	b) Participants must have clearly specified difficulty with mathematics	Participants do not have maths difficulties, or have general learning difficulties where maths difficulties not specified	Focus of the review is on whether manipulatives are effective with children with maths difficulties
4) Intervention	a) Must use concrete manipulatives as the focus of the intervention	Studies that do not focus on manipulative use, e.g. where it is combined with another approach	Focus of the review is the use of manipulatives; if combined with other interventions or strategies it will not be possible to address the review question
	b) Must use physical manipulatives	Studies that use computer-based manipulatives	Focus of the review is the use of concrete, physical manipulatives
	c) Children must	The use of manipulatives is not	This is important in the

	be taught how to use them	modelled or explained	effectiveness of the intervention (e.g. Carbonneau et al., 2013)
5) Date	Must be published before 15th Jan 2016	Studies published after 15th Jan 2016	To include all research published to date and allow time for analysis and write up

Figure 1: Process for study selection

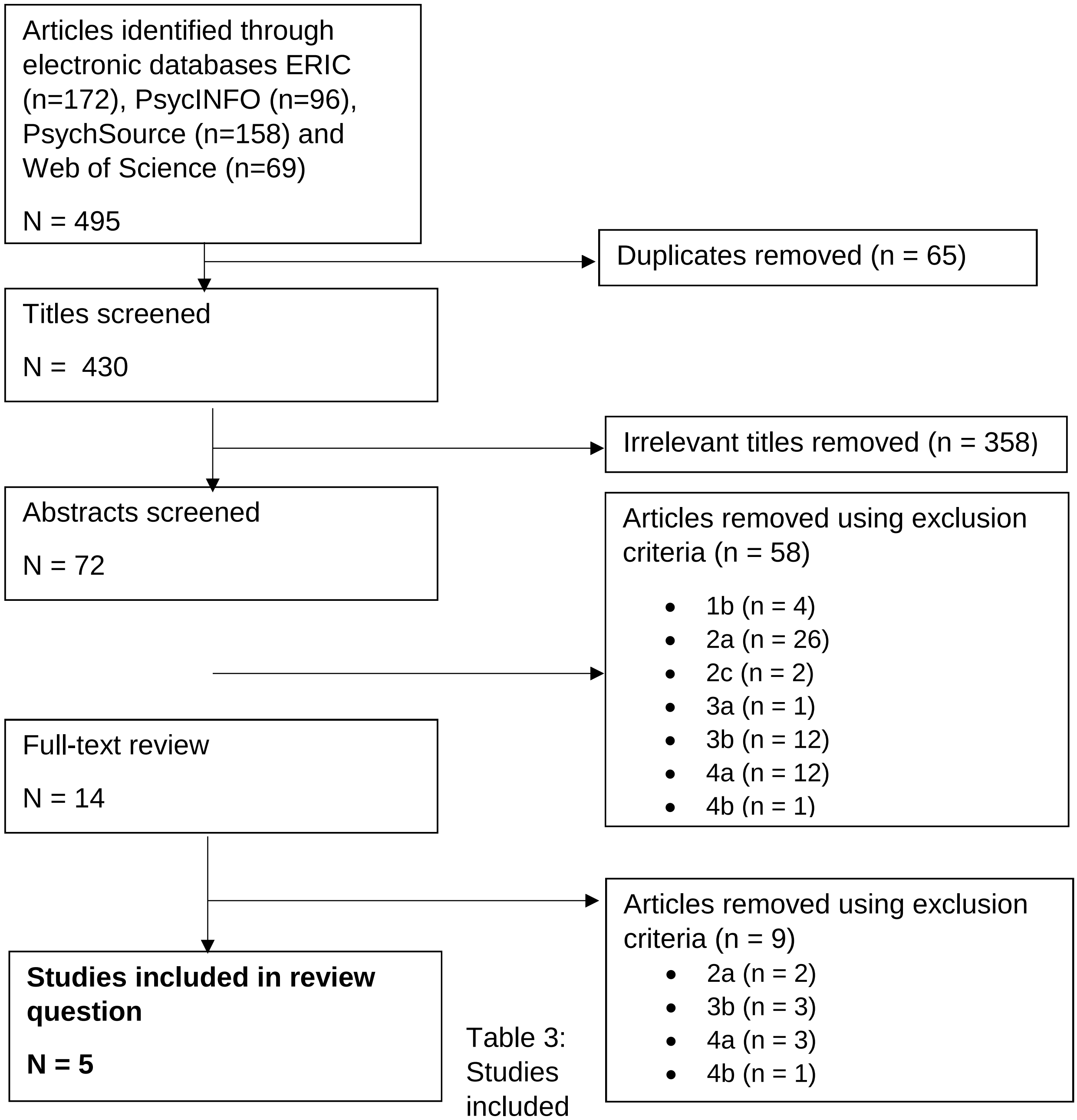

Table 3: Studies included in systematic review

Included studies
1. Butler, F. M., Miller, S. P., Crehan, K., Babbitt, B., & Pierce, T. (2003).

	Fraction Instruction for Students with Mathematics Disabilities: Comparing Two Teaching Sequences. *Learning Disabilities Research and Practice*, *18*(2), 99–111.
2.	Cass, M., Cates, D., & Smith, M. (2003). Effects of Manipulative Instruction on Solving Area and Perimeter Problems by Students with Learning Disabilities. *Learning Disabilities Research and Practice*, *18*(2), 112–120.
3.	Peterson, S., Mercer, C. D. & O'Shea, L. (1988). Teaching Learning Disabled Students Place Value Using the Concrete to Abstract Sequence. *Learning Disabilities Research, 4*(1), 52-56.
4.	Stroizer, S., Hinton, V., Flores, M., & Terry, L. (2015). An Investigation of the Effects of Model-Centered Instruction. *Education and Training in Autism and Developmental Disabilities*, *50*(2), 223–236.
5.	Witzel, B. S., Mercer, C. D., & Miller, M. D. (2003). Teaching Algebra to Students with Learning Difficulties: An Investigation of an Explicit Instruction Model. *Learning Disabilities Research and Practice*, *18*(2), 121–131.

Weight of Evidence

Key details of the studies reviewed are presented in Appendix 2. These were critically appraised using Gough's (2007) Weight of Evidence (WoE) framework which comprises of three judgements to examine the quality and relevance of each study.

WoE A evaluates methodological quality - the coherence and integrity of the evidence. Coding protocols were used to determine WoE A, completed examples of which can be found in Appendix 4. WoE B is a review-specific judgement on the relevance and appropriateness of the study's methodology to answering the review question. Finally WoE C addresses the relevance of the evidence to the review question. WoE D is formed by averaging these three judgements to assess the extent to which each study contributes to answering the review question. See Appendix 3 for further information on how the WoE judgements were calculated.

Table 4: Summary of WoE judgements

Study	WoE A Methodological quality	WoE B Methodological relevance	WoE C Relevance of evidence to the review question	WoE D Overall weight of evidence
Butler et al. (2003)	2 (Medium)	1 (Low)	2.6 (High)	1.9 (Medium)
Cass et al. (2003)	2.6 (High)	2 (Medium)	2.3 (Medium)	2.3 (Medium)
Peterson et al. (1988)	2 (Medium)	1 (Low)	1.3 (Low)	1.4 (Low)
Stroizer et al. (2015)	2.6 (High)	1 (Low)	2.6 (High)	2.1 (Medium)
Witzel et al. (2003)	2.25 (Medium)	3 (High)	3 (High)	2.75 (High)

The judgements are rated as follows:

- High: The study must receive an average score of 2.5 or above;
- Medium: The study must receive an average score between 1.5 and 2.4;
- Low: The study must receive an average score of 1.4 or below.

Critical appraisal

Participants

A total of 148 children between the ages of 8-15 participated in the studies reviewed, with numbers ranging from 3 for single case experimental designs (SCEDs) and 24-68 for the group designs. For the group designs, it is important to note that all three were underpowered; with a larger sample size they would have been more likely to detect any effect sizes present.

Participants' ages ranged from 8-16, although the majority included in this review (excluding Peterson et al., 1988 who reported a mean age of 10.4) were aged 12 and 13 (n = 95). All participants were sampled from schools in the USA; their race (reported by Cass et al., 2003 and Peterson et al., 1988), indicated a bias towards those from white ethnic backgrounds; and socio-economic status (SES) (reported by Peterson et al, 1988) showed a trend towards those from low SES backgrounds. These factors may impact on generalisability to UK school settings, however it must be noted that Peterson et al.'s (1988) study was considered low quality overall, so the sampling procedure should be interpreted with caution and may not be representative. Whilst the gender of participants was not reported in Witzel et al. (2003), the number of males (67%) outweighed females. Although the literature concerning gender differences in maths difficulties is mixed (Kikas, Peets, Palu & Afanasjev, 2009; Devine, Soltesz, Nobes, Goswami & Szucs, 2013) it does not support the presence of the large difference found in these studies.

All participants were considered to have some difficulty in maths, although criteria differed between studies, for example prior identification/diagnosis by professionals or performance on standardised tests. All studies used multiple methods of identification to triangulate information. The participants in some studies had additional difficulties such as a learning disability or Autism Spectrum Disorder (ASD) and this is likely due to it being relatively rare for a child to experience specific learning difficulties solely in maths (Chinn & Ashcroft, 1998). However, in order to answer the review question effectively, greater weighting was given in WoE C to studies which included participants solely with maths difficulties.

Design

This review examined studies with mixed designs. Single case design is "a rigorous, scientific methodology [which] has proven particularly relevant for defining educational practices at the level of the individual learner" (Horner, Carr, Halle, Mcgee, Odom & Wolery, 2005). These designs involve an effect replication across participants or behaviour, and more valid causal inferences are possible by staggering the intervention across one of these units (Kratochwill et al., 2010) as both studies did. However, the small sample sizes in SCEDs make generalisability difficult, and lack of control group means evidence is harder to attribute to the intervention. As such, randomised control designs were given a higher weighting in WoE B. Whilst all three studies featured active comparison groups with identifiable components, Butler et al. (2003) was given a lower rating due to its lack of a follow up group. As the overall aim of the intervention is to improve children's performance in particular areas of maths, this will only be possible if effects are maintained. Therefore, follow up data were deemed desirable. There was also variation in the use of follow up for the SCEDs, with Cass et al. (2003) given a higher WoE B weighting as it used a follow up across two further time points.

Randomisation was used in the group design studies to allocate participants to intervention or control groups which reduces the likelihood of selection bias as a threat to internal validity (Barker, Pistrang, Elliot & Barker, 2002). The SCEDs did not specify if participants were allocated to order of intervention delivery randomly (Cass et al., 2003) or if the order of behaviours in Stroizer et al. (2015) was randomly assigned. These SCEDs also screened participants on their relevant maths skills meaning they had baselines of zero. This may have led to effects that appear inflated and must be considered when interpreting their findings.

Measures

The studies used a range of measures directly related to their particular area of maths focus, which makes comparison challenging due to the lack of standardisation. Some, such as Butler et al. (2003) drew on established measures, while others were developed from existing educational materials (Cass et al., 2003) or by researchers specifically for the purpose of investigating progress in that particular area (Peterson et al., 1988). However, the studies and their measures were reviewed using the same criteria, and the reliability and validity of the measures varied considerably. For example, Peterson et al. (1988) scored zero for the measures criteria in WoE A as no reliability and validity criteria were reported, and thus its findings must be interpreted with extreme caution.

All studies with a follow-up design used the same measures as for post-test, with the exception of Cass et al. (2003) who used an additional practical long term generalisation test. Patton (2002) argues it is unwise to rely on one perspective, source or approach, yet only one study (Butler et al. 2003) included multi-source measures by employing a questionnaire on student attitudes. All other studies relied solely on a single measure of the skills trained in the intervention.

Additionally, no studies used a generic maths measure to investigate whether progress with specific areas of maths using manipulatives transferred to participants' overall maths performance. Had this data been available, it would have been valuable in contributing to the review question.

Intervention

A range of interventions using manipulatives as the fundamental basis have been presented. Their aims were the same – to improve skills in a specific area of maths – but the focus on different skills makes comparison more challenging. However, all of the studies except Cass et al. (2003) used a concrete-representational-abstract sequence of teaching, with all of these except Stroizer et al. (2015) using control groups who received teaching at only the representational-abstract or abstract level. This use of active comparisons enabled the component of concrete manipulatives to be identified; in quality studies significant differences between intervention and control groups could be attributed to the manipulative component as this was the only difference between groups.

Manipulative materials included beans, geoboards, cubes, sticks and markers, all of which are simple, low-cost resources. However, it important to note that all studies included a phase of teacher training and students were explicitly taught during the intervention how to use the manipulatives. This appears to be an important element of their success; this review shows that when teachers are trained in their use and explicitly model how students can use them, they can help improve maths skills. This is in line with Carbonneau et al. (2013) who argues that simply incorporating manipulatives into maths teaching may not be enough to increase achievement in maths, as high levels of guidance are associated with higher levels of student learning.

All studies assessed the fidelity of implementation and met acceptable levels for this as determined in WoE A. Therefore it can be inferred that, for manipulatives to produce meaningful change in maths skills, the teaching sequences described in the studies must be followed closely. However, there were differences in the manualisation of interventions; Peterson et al. (1988) was missing key information about the content of lessons while Witzel et al.'s (2003) intervention was clearly detailed. The high overall quality rating and significant effect sizes suggest this intervention is more favourable to use in schools.

There were also differences in the organisation of interventions. The group design studies (Butler et al. (2003) and Witzel et al. (2003)) taught with manipulatives as a whole class approach, whilst the other studies used 1:1 and 1:3 withdrawal teaching.

The differences in effect size, with SCEDs generally showing 'large' effects, could be due to the nature of delivery; the small group, individualised approach may have influenced the success of these participants. However, small group withdrawal is a method commonly used by schools to intervene with children experiencing difficulties with learning (see Dowker, 2004) and therefore can be considered generalisable. Additionally, there were differences between studies in the number and/or length of lessons participants received, ranging from 9x10-15min sessions (Peterson et al., 1988) to 19x50 min sessions (Witzel et al., 2003). Of the group design studies, Witzel et al. (2003) produced the greatest and most reliable effects, however its participants received nine more sessions than the participants in Butler et al.'s (2003) intervention. This will be an important consideration; it appears that more sessions may result in greater and more significant change.

Findings

As different designs were used, this review summarises study findings using Percentage of Non-overlapping Data (for SCEDs) and the Standardised Mean Difference of outcomes between intervention and control groups (for group designs). SMD was judged according to Cohen's (1988) effect size descriptors. Table 5 shows the indicators of effect size used to interpret the findings. Table 6 summarises the effect sizes for the five studies reviewed.

Table 5: Indicators of small, medium and large effect sizes according to method used

Type of effect size	Small (Questionable)	Medium (Effective)	Large (Highly effective)
Percentage of Non-Overlapping Data (PND) (Scruggs et al. 1986)	50-69%	70-89%	90-100%
Standardised Mean Difference (SMD) (Cohen, 1988)	0.20	0.50	0.80

Table 6: Effect sizes across studies

Study	Case/group	Outcome	Effect size type	Effect size	Effect size interpretation	Study quality rating (WoE D)
Butler et al. (2003)	Intervention group	Attitude	SMD	-0.02	No effect	1.9 (Medium)
		Area Fractions		0.13	No effect	
		Quantity Fractions		1.03*	Large*	
		Abstract Fractions		0.33	Small	
		Improper Fractions		0.32	Small	
		Word problems		0.17	No effect	
Cass et al. (2003)	Case 1	Perimeter	PND	92%	Highly effective	2.3 (Medium)
		Area		91%	Highly effective	
	Case 2	Perimeter		85%	Effective	
		Area		91%	Highly effective	
	Case 3	Perimeter		92%	Highly effective	
		Area		92%	Highly effective	
Peterson et al. (1988)	Intervention group	Place value acquisition	SMD	0.72*	Medium*	1.4 (Low)
		Place value generalisation		0.54	Medium	
Stroizer et al. (2015)	Case 1	Addition	SMD	100%	Highly effective	2.1 (Medium)
		Subtraction		91%	Highly effective	
		Multiplication		100%	Highly effective	
	Case 2	Addition		100%	Highly effective	
		Subtraction		90%	Highly effective	
		Multiplication		100%	Highly effective	
	Case 3	Addition		100%	Highly effective	
		Subtraction		90%	Highly effective	
		Multiplication		100%	Highly effective	
Witzel et al. (2003)	Intervention group	Algebra scores post	SMD	0.86*	Large*	2.75 (High)
		Algebra scores follow up		0.51*	Medium*	

*denotes statistical analyses significant to p<.05

Overall, Table 6 shows that manipulatives can be effective in improving the maths skills of children with maths difficulties. However, there is some variation in effect sizes so these findings should be considered in more detail. Both SCEDs showed large PNDs indicating high effects of intervention; the Stroizer et al. (2015) study in particular shows little variation between participants, although it seems subtraction skills were less successfully acquired. Whilst it did not use any follow up measures to see if these effects were maintained and hence received a lower weighting on WoE B, its results are in line with Cass et al. (2003) which also has an overall medium quality rating. The nature of the design meant control groups were not used, however they provide promising evidence that the small group approach used in these studies is effective in delivering manipulative intervention to improve the area, perimeter, addition, subtraction and multiplication skills of children with maths difficulties where they start from a baseline of zero.

The group design studies all show some statistically significant effects for manipulative intervention. It is interesting to reflect that in Butler et al. (2003), the intervention group showed bigger pre/post significant change for each fraction subtest than the control group, but the only significant difference between the groups was for Quantity Fractions. It is unclear whether this means manipulatives are particularly effective for this area of maths, although as with the other group designs, this study was underpowered and it is likely that any effect sizes present in other subtests would have been detected with a larger sample.

Peterson et al. (1988) found a medium effect of manipulative use on place value skill acquisition and no significant effect of generalisation of place value skill. However, these findings are considerably undermined by the overall low quality of the study which has issues with its use of measures, manualisation and implementation. Therefore, greater emphasis should be placed on the findings of sound studies such as Witzel et al. (2003), which received an overall high quality weighting. They report significant effects of manipulative use both at post-test and 3 week follow up, meaning it can be reliably inferred that manipulatives are effective in improving the maths skills of children with maths difficulties. The effect size post-intervention was large but only medium at 3 week follow up, meaning some level of learning was lost when intervention stopped. However, these results are very encouraging.

Conclusions and Recommendations

This review examined five studies to explore the effectiveness of manipulative interventions for children with mathematics difficulties. Overall, encouraging evidence was found in favour of the use of manipulatives in the areas of algebra, area and perimeter, addition, subtraction and multiplication and fractions. Additionally, they appear to facilitate change both on a whole class and small group/1:1 basis.

However, there were some key components present in all the studies which appear integral to the interventions' success. Intervention training was attended by all teachers and levels of fidelity were assessed. In practice, EPs could provide training to ensure teachers model the use of manipulatives effectively, with follow-up observations to support them in maintaining fidelity so manipulatives form an integral part of their teaching.

It was beyond the scope of this review to investigate the relationship between the number and/or length of sessions and impact on maths skill, however findings appear to indicate that more sessions may produce more effective change. The highest quality study (Witzel et al., 2003) found significant medium-large effects of intervention at both post-test and follow up using the most sessions. This tentative conclusion is in line with Dean and Kuhn (2007) who argue that in order for student-controlled strategies to be effective, students must engage in instruction over an extended period of time.

With the exception of Peterson et al. (1988), 88% of participants would be considered secondary school age (12-16), yet the use of manipulatives is only referred to in government guidance for Key Stages 1 and 2 (DfE, 2013), which covers primary school children aged 5-11. The studies reviewed show evidence that manipulatives are effective for use with older children to teach more complex mathematical concepts such as algebra (Witzel et al., 2003) and perimeter and area (Cass et al., 2003). They provide persuasive evidence that manipulatives should be extended beyond the UK primary school, a particularly pertinent finding when considering how to raise maths attainment so that fewer students leave secondary school without the functional maths skills required in society.

There are some limitations to consider. Firstly, none of the studies were UK based and research applying the principles to use in UK classroom settings is required to ensure findings generalise. To account for some of the methodological flaws in the existing literature, further research would need to use multi-measure methods with acceptable reliability and validity criteria along with follow up measures. Secondly, there is a lack of evidence about whether the gains in specific maths skills trained in the interventions generalise to overall maths performance. Fuson and Briars (1990) questioned whether there are problems with transfer when using manipulatives as students can fail to use the knowledge gained to solve written problems unless explicitly reminded. This would be a key consideration as teachers would want to see a measurable impact on students' maths attainment. Future research could be developed with an overall generalisation measure, for example a curriculum-based maths test administered pre- and post-intervention to assess whether this is the case.

Further developments could include extending and adapting the techniques used in the SCEDs for whole class use in order to assess whether the small group setting is important in producing the level of change, or if similar results can be gained when using the manipulatives in the classroom.

References

Amaya, M. M., Uttal, D. H., O'Doherty, K. D., Liu, L. L., & DeLoache, J. S. (2007). Two-digit sub-traction: Examining the link between concrete and abstract representations of knowledge. Cited in McNeil, N., & Jarvin, L. (2007). When theories don't add up: disentangling he manipulatives debate. *Theory Into Practice, 46*(4), 309–316.

Barker, C., Pistrang, N., Elliott, R., & Barker, C. (2002). *Research methods in clinical psychology*. Hoboken, N.J.: J. Wiley.

Bouck, E. C., Satsangi, R., Doughty, T. T. & Courtney, W. T. (2014). Virtual and Concrete Manipulative: A comparison of approaches for solving mathematics problems for students with autism spectrum disorder. *Journal of Autism Development Discord, 44,* 180-193.

Bruner, J. S. (1964). The course of cognitive growth. *American Psychologist, 19*(1), 1–15.

Bryant, D. P., Bryant, B. R., Shin, M. & Pfannenstiel, K. H. (2014). Learning disabilities: Mathematics characteristics and instructional exemplars. In S. Chinn (Ed.), *The Routledge international handbook of dyscalculia and mathematical learning difficulties* (pp. 243-256). London: Routledge.

Butler, F. M., Miller, S. P., Crehan, K., Babbitt, B., & Pierce, T. (2003). Fraction Instruction for Students with Mathematics Disabilities: Comparing Two Teaching Sequences. *Learning Disabilities Research and Practice*, *18*(2), 99–111.

Carbonneau, K. J., Marley, S. C., Selig, J. P. (2013). A meta-analysis of the efficacy of teaching mathematics with concrete manipulatives. *Journal of Educational Psychology, 105*(2), 380-400.

Cass, M., Cates, D., & Smith, M. (2003). Effects of Manipulative Instruction on Solving Area and Perimeter Problems by Students with Learning Disabilities. *Learning Disabilities Research and Practice*, *18*(2), 112–120.

Clements, D. H. (1999). Concrete manipulatives, concrete ideas. *Contemporary Issues in Early Childhood, 1,* 45-60.

Chinn, S. (2014). *The Routledge international handbook of dyscalculia and mathematical learning difficulties*. London: Routledge.

Chinn, S. & Ashcroft, J. R. (1998). *Mathematics for Dyslexics.* London: Whurr.

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

Dean, D., & Kuhn, D. (2007). Direct instruction vs. discovery: The long view. *Science Education, 91*(3), 384–397.

Department for Education (DfE) (2013). *Mathematics programmes of study: Key stages 1 and 2: National Curriculum in England.* Reference: DFE-00180-2013

Devine, A., Soltesz, F., Nobes, A., Goswami, U & Szucs, D. (2013). Gender differences in developmental dyscalculia depend on diagnostic criteria. *Learning and Instruction, 27,* 31-39.

Domino, J. (2010). *The effects of physical manipulatives on achievement in mathematics in grades K-6: A meta-analysis* (Doctoral dissertation). Retrieved from ProQuest. (UMI 3423451)

Dowker, A. (2004). *What works for children with mathematical difficulties? (*Research Report RR554). Nottingham: DfES Publications.

Fletcher, D., Boon, R. T., & Cihak, D. F. (2010). Effects of the TOUCHMATH program compared to a number line strategy to teach addition facts to middle school students with moderate intellectual disabilities. Education and Training in Developmental Disabilities, 45, 449–458.

Flores, M. M., Hinton, V. M., & Schweck, K. B. (2014). Teaching Multiplication with Regrouping to Students with Learning Disabilities, 29(4), 171–183.

Fuson, K. C., & Briars, D. J. (1990). Using a base-ten blocks learning/teaching approach for first-and second-grade place-value and multidigit addition and subtraction. *Journal for Research in Mathematics Education, 21*, 180–206.

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. H., Sackett, D. L., Sinclair, J. C., Hayward, R., Cook, D. J. & Cook, R. J. (1995). Users' guides to the medical literature IX. A method for grading health care recommendations. *Journal of the American Medical Association, 274*(22), 1800-1804.

Hartshorn, R. & Boren, S. (1990). *Experiential learning of mathematics: Using manipulatives.* Charleston, WV: ERIC Clearinghouse on Rural Education and Small Schools.

Holmes, W. & Dowker, A. (2013). Catch Up Numeracy: a targeted intervention for children who are low-attaining in mathematics. *Research in Mathematics Education, 15*(3), 249-265.

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.

Jordan, L., Miller, D. & Mercer, C. (1999). The Effects of Concrete to Semiconcrete to Abstract Instruction in the Acquisition and Retention of Fraction Concepts and Skills. *Learning Disabilities: A Multidisciplinary Journal, 9*(3), 115-122.

Kikas, E., Peets, K., Palu, A. & Afanasjev, J. (2009). The role of individual and contextual factors in the development of maths skills. *Educational Psychology, 29*(5), 541-560.

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

Kratochowill, T.R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., Shadish, W. R. (2010). Single-case design technical documentation. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/documentsum.aspx?sid=229

Maccini, P., & Hughes, C. A. (2000). Effects of a problem-solving strategy on the introductory algebra performance of secondary students with learning disabilities. *Learning Disabilities Research & Practice*, *15*, 10–21.

Martin, T., & Schwartz, D. L. (2005). Physically distributed learning: Adapting and reinterpreting physical environments in the development of fraction concepts. *Cognitive Science, 29*, 587–625.

McNeil, N., & Jarvin, L. (2007). When theories don't add up: disentangling he manipulatives debate. *Theory Into Practice, 46*(4), 309–316.

Moyer, P. S. (2001). Are we having fun yet? How teachers use manipulatives to teach mathematics. *Educational Studies in Mathematics, 47*, 175–197.

National Council of Teachers of Mathematics (NCTM). (2000). *Principles and standards for school mathematics.* Reston, VA: National Council of Teachers of Mathematics.

National Numeracy (2014). *What is the issue? Low levels of numeracy are a long-term problem for the UK.* Retrieved from: http://www.nationalnumeracy.org.uk/what-issue.

Patton, M.Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.

Peterson, S., Mercer, C. D. & O'Shea, L. (1988). Teaching Learning Disabled Students Place Value Using the Concrete to Abstract Sequence. *Learning Disabilities Research, 4*(1), 52-56.

Piaget, J. (1965). *The child's conception of number.* New York: W. W. Norton & Company.

Rashid, S., & Brooks, G. (2010). *The levels of attainment in literacy and numeracy of 13- to 19-year-olds in England, 1948-2009.* London: NRDC.

Reeve, R. A. & Gray, S. (2014). Number difficulties in young children. In S. Chinn (Ed.), *The Routledge international handbook of dyscalculia and mathematical learning difficulties* (pp. 44-59). London: Routledge.

Rittle-Johnson, B., & Koedinger,K. R. (2005). Designing knowledge scaffolds to support mathematical problem solving. *Cognition and Instruction, 23*, 313–349.

Scruggs, T. E., Mastropieri, M. A., Cook, S. B., & Escobar, C. (1986). Early interventions for children with conduct disorders: A quantitative synthesis of single-subject research. Behavior Disorders, 260-271.

Stellingwerf, B. P., & Lieshout, E. C. D. M. (1999). Manipulatives and number sentences in computer aided arithmetic word problem solving. *Instructional Science*, *27*(6), 459–476.

Sternberg, R. J., & Grigorenko, E. L. (2004). Successful intelligence in the classroom. *Theory Into Practice*, 43, 274–280.

Stroizer, S., Hinton, V., Flores, M., & Terry, L. (2015). An Investigation of the Effects of Model-Centered Instruction. *Education and Training in Autism and Developmental Disabilities*, *50*(2), 223–236.

Thompson, P. W. (1992). Notations, conventions, and constraints: Contributions to effective uses of concrete materials in elementary mathematics. *Journal for Research in Mathematics Education, 23,* 123–147.

Underhill, B. (1981). *Teaching elementary school mathematics* (3rd ed.). Columbus, OH: Merril.

Witzel, B. S. (2005). Using CRA to Teach Algebra to Students with Math Difficulties in Inclusive Settings. *Learning Disabilities: A Contemporary Journal*, *3*(2), 49-60.

Witzel, B. S., Mercer, C. D., & Miller, M. D. (2003). Teaching Algebra to Students with Learning Difficulties: An Investigation of an Explicit Instruction Model. *Learning Disabilities Research and Practice*, *18*(2), 121–131.

Appendix 1

Studies excluded at full text review

Excluded paper	Rationale for exclusion
Flores, M. M., Hinton, V. M., & Schweck, K. B. (2014). Teaching Multiplication with Regrouping to Students with Learning Disabilities, 29(4), 171–183.	Exclusion criteria: 4a Intervention also used a Strategic Instruction Model
Jimenez, B. A., Browder, D. M., & Courtade, G. R. (2008). Teaching an algebraic equation to high school students with moderate developmental disabilities. *Education and Training in Developmental Disabilities*, *43*(2), 266–274.	Exclusion criteria: 3b Participants had 'moderate developmental disabilities' but not clearly specified maths difficulties
Jordan, L., Miller, D. & Mercer, C. (1999). The Effects of Concrete to Semiconcrete to Abstract Instruction in the Acquisition and Retention of Fraction Concepts and Skills. *Learning Disabilities: A Multidisciplinary Journal, 9*(3), 115-122.	Exclusion criteria: 3b Some participants had 'disabilities' but study didn't specify their maths abilities
Maccini, P., & Ruhl, K. L. (2000). Effects of a Graduated Instructional Sequence on the Algebraic Subtraction of Integers by Secondary Students with Learning Disabilities. *Education and Treatment of Children*, *23*(4), 465–489.	Exclusion criteria: 4a Study also used general problem solving and self-monitoring strategies
Mancl, D. B., Miller, S. P., & Kennedy, M. (2012). Using the Concrete-Representational-Abstract Sequence with Integrated Strategy Instruction to Teach Subtraction with Regrouping to Students with Learning Disabilities. *Learning Disabilities Research & Practice*, *27*(4), 152–166.	Exclusion criteria: 4a Intervention also used an Integrated Strategy Instruction approach
Morgan, P. L., Farkas, G., & Maczuga, S. (2014). Which instructional practices most help first-grade students with and without mathematics difficulties? *Educational Evaluation and Policy Analysis*, *36*(3), 1–22.	Exclusion criteria: 2a Study analyses pre-existing longitudinal data
Pasnak, R., Hansbarger, A., Dodson, S. L., Hart, J. B., & Blaha, J. (1996). Differential results of instruction at the preoperational/concrete operational transition. *Psychology in the Schools, 33*(1), 70–83.	Exclusion criteria: 3b Participants did not have clearly specified maths difficulties
Stellingwerf, B. P., & Lieshout, E. C. D. M. (1999). Manipulatives and number sentences in computer aided arithmetic word problem solving. *Instructional Science*, *27*(6), 459–476.	Exclusion criteria: 4b Study used computer programme manipulatives
Witzel, B. S. (2005). Using CRA to Teach Algebra to Students with Math Difficulties in Inclusive Settings. *Learning Disabilities: A Contemporary Journal*, *3*(2), 49-60.	Exclusion criteria: 2a Reports data previously published in Witzel (2003)

Appendix 2
Summary of included studies

Author	Sample (participants)	Design	Intervention	Outcome measures	Findings
Butler et al. (2003)	50 (intervention group = 26) Intervention group: 15 males, 11 females Age: 11-15 Control group: 12 males, 12 females Age: 11-14 Maths difficulty identified using: • Prior identification of specific learning disabilities in maths • Receiving maths instruction in resource room settings	Randomised control trial	Manipulative use for fraction concepts and procedures Manipulatives used: fraction circles, dried beans, student-made fraction squares 10x 45 minute lessons; daily Intervention/control groups formed of two separate classes taught by same teacher to control for teacher effects. 2 teachers participated Intervention group received concrete-representational-abstract lessons whilst control received only representational-abstract lessons, i.e. without use of manipulatives Setting: intervention delivered by teacher in whole class format as part of normal maths teaching in resource room setting in US public school	Subtests taken from Brigance Comprehensive Inventory of Basic Skills-Revised (*) and researcher-designed measures. CIBS-R has high reliability and validity. Researcher-designed measures had high correlations with subtests taken from CIBS-R Pre/post measures: • 5 subtests to measure performance on fraction concepts, procedures and application: o Quantity Fractions * o Area Fractions * o Abstract Fractions * o Word Problems o Improper Fractions • Student attitude towards mathematics questionnaire – 3 point Likert scale Follow up data not collected	Both groups showed statistically significant improvement on all measures except the intervention group on the Area Fractions subtest Although not statistically significant, the mean scores for the intervention group were higher than those of the control. Effect sizes were larger for the intervention group on all subtests except Area Fractions Post hoc tests showed a significant difference only between the control and intervention groups for Quantity Fractions Both groups' attitudes to maths instruction improved statistically significantly

Author	Sample (participants)	Design	Intervention	Outcome measures	Findings
Cass et al. (2003)	3 2 males, 1 female Age: 13-16 Maths difficulty measured using: • Significant discrepancy between intelligence and maths achievement • Area and perimeter score of <40% • Identified by teacher and caregiver as requiring area and perimeter problem solving skills	Multiple baseline across participants and behaviour design	Manipulative use for teaching perimeter and area problem solving Manipulatives used: geoboards, rubber bands, 25ft measuring tape Sessions were 15-20 minutes daily. Number of intervention sessions each child received depended on their performance; criterion to move onto next behaviour then maintenance check phase was 80% or more problems correct on 3 consecutive days Students were taught using a sequence of teachers modelling the use of manipulatives for each skill, prompted/guided practice with them, modelling how to solve a question and independent practice with them on questions which served as the daily test Setting: intervention delivered 1:1 in resource room setting by teacher certified in special	2 behaviours probed: • Perimeter problem solving ability • Area problem solving ability Problems used to probe behaviours selected from three US maths textbooks Area and perimeter problem solving skills deemed acquired when: • Each student solved 80% or more problems correct without prompting • On three consecutive days Maintenance checks performed 2x per week for 3 weeks once child had reached criterion for both behaviours After two week Christmas break, long term check. Also generalisation measure applying skills to calculate floor and window sizes in doll's house	0 students were able to solve area and perimeter problems correctly during baseline. All achieved the performance criterion and maintained these skills to the end of follow up testing after a Christmas break On average, the students took 5.5 days to acquire each skill. Perimeter took an average of 6 days to acquire and area took an average of 5 days

			education in US public school		
Author	Sample (participants)	Design	Intervention	Outcome measures	Findings
Peterson et al. (1988)	24 (intervention group = 12) 20 males, 4 females Intervention group: 10 males, 2 females Age = 9-13 Control group: 10 males, 2 females Age = 8-12 Maths difficulty measured using: • Maths achievement on WRAT-R • Maths instruction in special education classrooms	Randomised control trial	Manipulative use for teaching place value skills of identifying ones and tens in 2 digit numbers Manipulatives used: unifix cubes, place value sticks (lolly sticks) and place value strips 9 x 10-15 minute lessons; daily Intervention group received 3 concrete > 3 semi-concrete > 3 abstract lessons in sequence while control group received 9 lessons at abstract level i.e. without manipulatives, just numbers Setting: intervention delivered by teacher and researcher in groups of three in resource room classrooms in a US public school	Performance on researcher-created questions on place value in 2 digit numbers. 2 aspects: acquisition and generalisation Generalisation questions tested a different stimulus: place value in 3 and 4 digit numbers Maintenance measured one week after end of intervention when instruction ceased Retention measured three weeks after end of intervention	Significant main effect for instructional treatment found – those in intervention group using manipulatives performed better than those in control on acquisition No main effect found for generalisation – both groups performed similarly when tested on untaught skill of identifying ones and tens in 3 and 4 digit numbers
Stroitzer et al. (2015)	3 3 males Age: 8-10	Multiple baseline across behaviours design	Manipulative use for teaching 3 maths concepts to children with ASD. Followed concrete-representational-abstract sequence	3 behaviours probed: • Addition with regrouping • Subtraction with regrouping	2 students started at baselines of 0 for all 3 behaviours and the other student had a baseline of 0 for 2 out of 3 of the

Participants had ASD diagnosis Maths difficulty measured using: • Difficulty with maths achievement as assessed using Keymath III • Received special education services in the areas of maths	Manipulatives used: base-ten foam blocks, selection of objects 20 x 20-60 minute sessions (20 mins per behaviour); daily Students were taught using a concrete > representational > abstract sequence. Teacher modelling of and independent practice with manipulatives used in concrete phase. When students completed independent practice problems with 80% accuracy, moved on to representational phase. Abstract instruction started in the 8th lesson for each behaviour and continued until the end of the programme. Setting: intervention delivered by teacher in small group in US summer school. Students began instruction together, but when reached criterion received instruction for next behaviour and got additional teaching so size fluctuated between 1-3	• Multiplication facts zero to five Assessed using researcher-designed probes rated for content validity and difficulty Follow up/maintenance/ generalisation data not collected	behaviours. The criterion for phase change was when 6 out of 9 problems correct was achieved. All students achieved criterion. Tau-U was calculated for each student and behaviour. The overall effect sizes for all three students across behaviours was .97 Demonstrated a functional relation between the CRA instruction and the three behaviours

Author	Sample (participants)	Design	Intervention	Outcome measures	Findings
Witzel et al. (2003)	358 (study focused on a subset of 68 children; 34 matched pairs) Genders not specified Intervention group: 34 participants Age: 11-14 Control group: 34 participants Age: 11-14 Maths difficulty measured using: • Identified as having a learning disability by school services (1.5 SD discrepancy between ability and achievement with reference specifically to maths • Performed below average in classroom according to	Randomised control trial	Manipulative use for teaching algebraic transformation equations Manipulatives used: minus/plus and coefficient markers, large sticks, equals lines and small sticks 19x 50 minute lessons; daily Intervention/control groups formed of two separate classes taught by same teacher to control for teacher effects. 4 schools and 10 teachers participated Intervention group received concrete-representational-abstract lesson sequence, while control received abstract only lessons = difference between groups was use of manipulatives Setting: intervention delivered by teachers in whole class format as part of normal maths teaching in US public	Pre/post performance on researcher-designed and teacher-validated algebra test 70 questions developed from curriculum material and distributed to 4 algebra teachers who accepted/ rejected/improved them. Revised 63 questions given to 32 students who had completed their pre-algebra course and responses used to calculate item difficulty. 27 items with a medium difficulty level were selected for the pre/post measure Pre-test administered one week prior to intervention, post-test on day of completion and follow-up three weeks after intervention ended	Both groups showed statistically significant improvement in their responses to solving algebraic equations from pre to post test Both groups also improved significantly from pre to follow up test. However, there was no reliable change from post-test to follow up The interaction between test occasion and treatment condition yielded a significant difference; the intervention group outperformed the control group on post-test and follow-up

maths teacher • Scored below 50th percentile on most recent statewide achievement test	schools

Appendix 3
Weighting of studies

WoE A: Methodological quality

WoE A was assessed using an adapted version of Kratochwill's (2003) coding protocol for group designs and Horner et al.'s (2005) protocol for single case designs. These protocols enabled objective judgements to be made on the methodological quality of each study.

Judgements using the Kratochwill (2003) protocol are weighted on four of the criteria: measurement, comparison group, fidelity and follow up. Where there is more than one descriptor for each weighting, the study had to meet every descriptor to be given that weighting. Justification for scores given are based on the following guidance:

Table 1: WoE A criteria for Measurement:

Weighting	Descriptor
High 3	Used measures that produce reliable scores of at least .70, for the majority of primary outcomes Data was collected using multiple methods, and collected from multiple sources Validity is reported
Medium 2	Used measures that produce reliable scores of at least .70 Data was collected using multiple methods, and/or collected from multiple sources A case for validity does not need to be presented
Low 1	Used measures that produce reliable scores of at least .50 Data may have been collected either using multiple methods and/or from multiple sources however, this is not required. A case for validity does not need to be presented

Table 2: WoE A criteria for Comparison group:

Weighting	Descriptor
High 3	Uses at least one type of "active" comparison group Initial group equivalency must be established Evidence that change agents were counterbalanced Equivalent mortality and low attrition at post, and if applicable, at follow-up
Medium 2	Uses at least a "no intervention group" type of comparison There is evidence for at least two of the following: counterbalancing of change agents, group equivalence established, or equivalent mortality with low attrition If equivalent mortality is not demonstrated, an intent-to intervene

	analysis is conducted
Low 1	Uses a comparison group At least one of the following is present: counterbalancing of change agents, group equivalence established, or equivalent mortality with low attrition If equivalent mortality is not demonstrated, an intent-to-intervene analysis is conducted

Table 3: WoE A criteria for Implementation fidelity:

Weighting	Descriptor
High 3	Demonstrates strong evidence of acceptable adherence Evidence of fidelity is measured through at least two of the following: ongoing supervision/consultation, coding sessions, or audio/video tapes, and use of a manual The “manual” is either written materials involving a detailed account of the exact procedures and the sequence in which they are to be used or formal training session detailing exact procedures and sequence
Medium 2	Demonstrates evidence of acceptable adherence Evidence of fidelity is measured through at least one of the following: ongoing supervision/consultation, coding sessions, or audio/video tapes, and use of a manual. The “manual” is either written materials involving an overview of broad principles and a description of the intervention phases, or formal/informal training session involving an overview of broad principles and a description of the intervention phases
Low 1	Demonstrates evidence of acceptable adherence measured through at least one of the above criteria or use of a manual

Table 4: WoE A criteria for Follow up:

Weighting	Descriptor
High 3	Conducted follow up assessments over multiple intervals with all participants that were included in the original sample Uses similar measures used to analyse data from primary or secondary outcomes
Medium 2	Conducted follow up assessments at least once with the majority of participants that were included in the original sample Similar measures used to analyse data from primary or secondary outcomes
Low 1	Conducted follow up assessments at least once with some participants from the original sample

For the single case design studies, Horner et al.'s (2005) protocol is used. Ratings are given according to the 6 criteria detailed in Table 6 on a scale of 3, 2 or 1 depending on how closely the study meets the sub-criteria for each area. Completed coding protocols can be found in Appendix 4 along with justification for these ratings.

To get an overall score for WoE A, these scores are averaged over the criteria (of 4 for group designs and 6 for single case designs). To receive a weighting of:

- High: The study must receive an average score of 2.5 or above;
- Medium: The study must receive an average score between 1.5 and 2.4;
- Low: The study must receive an average score of 1.4 or below.

Table 5: WoE A ratings for group design studies

Study	Measurement	Comparison group	Implementation fidelity	Follow up	WoE A
Butler et al. (2003)	2	3	3	0	2 (Medium)
Peterson et al. (1988)	0	3	2	3	2 (Medium)
Witzel et al. (2003)	1	3	3	2	2.25 (Medium)

Table 6: WoE A ratings for single case design studies

Study	Participants and settings	Dependent variable	Independent variable	Baseline	Experimental control/ internal validity	Social validity	WoE A
Cass et al. (2003)	3	2	3	3	2	3	2.6 (High)
Stroitzer et al. (2015)	2	3	3	3	2	3	2.6 (High)

WoE B: Methodological relevance

WoE B considers whether the design was suitable for evaluating the effectiveness of manipulatives on the maths performance of children with maths difficulties. Criteria are based on evidence hierarchies (e.g. Guyatt et al., 1995) in which randomised controlled trials are deemed to provide evidence of higher quality as they control for threats to internal validity, while those with high threats to internal validity (e.g. no control group) and single case designs are placed lower down.

The following descriptors were considered methodologically relevant to the review question as an active comparison group where only one component is changed – i.e. the use of manipulatives – will demonstrate whether the intervention had an effect because of their use. Pre and post measures are necessary to compare performance, and follow-up measures are deemed methodologically relevant as they are essential to assess whether any improvements made are maintained after the intervention ceases. This is a crucial factor schools need to consider when choosing interventions. WoE B was assessed as follows, where the study had to meet all of the descriptors within its weighting to be awarded that judgement:

Weighting	Descriptor
High 3	• Group design o Randomisation to groups o Active comparison group o Outcome data collected pre and post intervention o Follow up data
Medium 2	• Group design o May be random allocation o Control group (wait-list or no intervention) o Outcome data collected pre and post intervention o Follow up data • Single case design o Multiple baseline across participants or at least three attempts to demonstrate intervention effect o Generalisation/maintenance and follow up with at least 3 data points
Low 1	• Group design o No control group o No pre and post outcome data o No follow up data • Single case design o Intervention effect not demonstrated 3 times o Generalisation/maintenance data not included/less than 3 data points o No follow up

WoE C: Relevance of evidence to the review question (topic relevance)

WoE C considers the extent to which the topic of the study and its findings contribute to answering the review question. WoE C criteria were considered on three factors. Firstly, as the manipulative interventions included in the review were all different, it was considered important to the review question that the intervention was fully manualised to assess to what extent manipulatives were, how they were used and to facilitate the replication of the intervention. Secondly, this review examines whether manipulatives are effective for use with children with maths difficulties. Therefore, studies that assess maths difficulties on a number of criteria and exist alone e.g. without comorbidity with general learning difficulties are given greater weighting as they are more able to answer the review question. Finally, this review considers whether manipulatives are effective for those with maths difficulties in the context of a school setting. Therefore, it is important that interventions described in the studies are delivered by a teacher to enhance the generalisability of findings to their use in the classroom. The weightings for WoE C are considered as follows, where the study had to meet all of the descriptors within its weighting to be awarded that judgement:

Manualisation

Weighting	Descriptor
High 3	• Intervention described in specific detail; replicable
Medium 2	• Intervention described in fair detail
Low 1	• Intervention described without enough information to understand the intervention or attempt to replicate it

Maths difficulties

Weighting	Descriptor
High 3	• Maths difficulty assessed using 3 or more criteria • Maths difficulties alone
Medium 2	• Maths difficulty assessed using 2 or more criteria • Maths difficulties in presence of other specific needs, e.g. ASD
Low 1	• Maths difficulty assessed using 2 or more criteria • Maths difficulties in presence of other general needs, e.g. previously diagnosed learning difficulties/disabilities

Implementation

Weighting	Descriptor
High 3	• Implemented by teachers
Medium 2	• Implemented by researchers and teachers
Low 1	• Implemented by researchers

Table 7: WoE C ratings

Study	Manualisation	Maths difficulties	Implementation	WoE C
Butler et al. (2003)	3	2	3	2.6 (High)
Cass et al. (2003)	3	1	3	2.3 (Medium)
Peterson et al. (1988)	1	1	2	1.3 (Low)
Stroitzer et al. (2015)	3	2	3	2.6 (High)
Witzel et al. (2003)	3	3	3	3 (High)

Appendix 4
Coding Protocols

Coding protocol for Methodological Quality Weight of Evidence A (WoE A): Group based design

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

Coding Protocol: Group-Based Design

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**: 28/01/2016

Full Study Reference in APA format: Butler, F.M, Miller, S.P., Crehan, K., Babbitt, B. & Pierce, T. (2003) Fraction Instruction for Students with Mathematics Disabilities: Comparing Two Teaching Sequences. *Learning Disabilities Research and Practice, 18*(2), 99-111.

Intervention Name (description from study): Teaching equivalent fraction concepts using the concrete-representational-abstract (CRA) sequence

Study ID Number (Unique Identifier): 1

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: MANCOVA
_level: .05
ES: medium (.5) Medium has been selected as the authors have not stated which effect size they were looking for
N required: 64

B4. Total size of sample (start of the study): 50 ☐ N ☐
B5. Intervention group sample size: 26 ☐ ☐
N
B6. Control group sample size: 24 ☐ ☐
N

~~**For studies using qualitative research methods, code B7 and B8**~~ This section has been removed as the studies were not qualitative

~~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**~~ This section has been removed as it is not relevant to the question of methodological quality; outcomes are examined separately in this review

~~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: 1
N

E4. Number of components linked to primary outcomes: 1
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_No follow up assessment____

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____

Specify rationale for sample size: _No____

A1.1 Inclusion/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~~ This section has been removed as participant information is reported elsewhere in this review

~~Participants from Treatment Group~~	~~Grade/age~~	~~Gender~~	~~Ethnicity or Multi-ethnic~~	~~Ethnic Identity~~	~~Race(s)~~	~~Acculturation~~	~~Pri-mary Lan-guage~~	~~SES~~	~~Family Struc-ture~~	~~Locale~~	~~Disability~~	~~Functional Descriptors~~
☐ ~~Child/Student~~ ☐ ~~Parent/caregiver~~ ☐ ~~Teacher~~ ☐ ~~School~~ ☐ ~~Other~~												
☐ ~~Child/Student~~ ☐ ~~Parent/caregiver~~ ☐ ~~Teacher~~ ☐ ~~School~~ ☐ ~~Other~~												
☐ ~~Child/Student~~ ☐ ~~Parent/caregiver~~ ☐ ~~Teacher~~ ☐ ~~School~~ ☐ ~~Other~~												
☐ ~~Child/Student~~ ☐ ~~Parent/caregiver~~ ☐ ~~Teacher~~ ☐ ~~School~~ ☐ ~~Other~~												

~~Participants from Control Group~~	~~Grade/age~~	~~Gender~~	~~Ethnicity or Multi-ethnic~~	~~Ethnic Identity~~	~~Race(s)~~	~~Acculturation~~	~~Pri-mary Lan-guage~~	~~SES~~	~~Family Struc-ture~~	~~Locale~~	~~Disability~~	~~Functional Descriptors~~
☐ ~~Child/Student~~ ☐ ~~Parent/caregiver~~ ☐ ~~Teacher~~ ☐ ~~School~~ ☐ ~~Other~~												
☐ ~~Child/Student~~ ☐ ~~Parent/caregiver~~ ☐ ~~Teacher~~ ☐ ~~School~~ ☐ ~~Other~~												
~~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: ____

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

Specify: ______________________________

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	Responses on attitude questionnaire showed statistically significant increase in student attitude towards maths instruction post-intervention.	☐Participants reported benefiting overall from the intervention ☐Participants reported not benefiting overall from the intervention
☐Child/Student ☐Parent/caregiver ☐Teacher ☐School ☐Other		☐Participants reported benefiting overall from the intervention ☐Participants reported not benefiting overall from the intervention
☐Child/Student ☐Parent/caregiver ☐Teacher ☐School ☐Other		☐Participants reported benefiting overall from the intervention ☐Participants reported not benefiting 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 2
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 45 mins
N

C2.2 frequency of intervention session daily
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
E9. ☐ Unknown/insufficient information provided

~~**F. Characteristics of the Intervener**~~

Section removed as no information given on 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~~
~~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 __1__________

Average length of training __3 hours______

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	2	Promising
Comparison Group	3	Strong
~~Primary/Secondary Outcomes are Statistically Significant~~		
Educational/clinical significance	2	Promising
Identifiable Components	2	Promising
Implementation Fidelity	3	Strong
Replication	1	Weak
Site of Implementation	3	Strong
Follow Up Assessment Conducted	0	No evidence

Numerical descriptors for these criteria not relevant to WoE A

~~**Descriptive or Supplemental Criteria**~~		
~~External validity indicators~~		
~~Length of Intervention~~		
~~Intensity/dosage~~		
~~Dosage Response~~		
~~Program Implementer~~		
~~Characteristics of the Intervener~~		
~~Intervention Style/Orientation~~		
~~Cost Analysis Data Provided~~		
~~Training and Support Resources~~		
~~Feasibility~~		

Appendix 4
Coding Protocols

Coding protocol for Methodological Quality Weight of Evidence A (WoE A):
Single case design

Single case design coding protocol adapted from: 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.

Coding Protocol

Study ID number: 2
Research design: Single case experimental – multiple baseline across participants

Name of coder: **Date:** 19/01/16

Full study reference in proper format: Cass, M., Cates, D., & Smith, M. (2003). Effects of Manipulative Instruction on Solving Area and Perimeter Problems by Students with Learning Disabilities. *Learning Disabilities Research and Practice, 18(*2), 112–120.

Type of publication:
☐Book/monograph
☒Journal article
☐Book chapter
☐Other (specify):

1. Description of Participants and Settings
Participants are described with sufficient detail to allow others to select individuals with similar characteristics (e.g. age, gender, disability, diagnosis)
☒Yes *3 participants, all characteristics described*
☐No
☐N/A
☐Unable to code

The process for selecting participants is described with replicable precision
☒Yes *4 criteria for participation clearly specified*
☐No
☐N/A
☐Unable to code

Critical features of the physical setting are described with sufficient precision to allow replication
☒Yes *1:1 basis in resource room setting*
☐No

☐N/A
☐Unable to code

Overall Rating of Evidence: ☒3 ☐2 ☐1 ☐0
This score was given as the study fulfilled all criteria

2. Dependent variable
Dependent variables are described with operational precision
☒Yes *Defined as performance on area and perimeter problems either correct or incorrect. Performance criterion defined as 80%+ problems solved correctly on 3+ days*
☐No
☐N/A
☐Unable to code

Each dependent variable is measured with a procedure that generates a quantifiable index
☒Yes *Problems were recorded correct/incorrect and errors recorded and classified into 4 groups*
☐No
☐N/A
☐Unable to code

Measurement of the dependent variable is valid and described with replicable precision
☐Yes
☒No *Area and perimeter problems randomly selected from 3 textbooks, but which ones or how many not specified*
☐N/A
☐Unable to code

Dependent variables are measured repeatedly over time
☒Yes *DVs measured at baseline, through intervention, maintenance phase and a follow up*
☐No
☐N/A
☐Unable to code

Data are collected on the reliability or interobserver agreement associated with each dependent variable, and IOA levels meet minimal standards (e.g. IOA = 80%; Kappa = 60%)
☒Yes *Observer scored students' solutions during 25% of treatment phase; 100% IO agreement*
☐No
☐N/A
☐Unable to code

Overall Rating of Evidence: ☐3 ☒2 ☐1 ☐0
This rating was given as there was an operationalised definition given for the DVs, they were quantifiable and 3+ data points were recorded. There was also high IOA of 100%. However, measurement of the DVs is not described with enough precision to replicate.

2. Independent variable

Independent variable is described with replicable precision

☒Yes *Manipulatives and materials used are clearly specified; sequence and content of lessons described in sufficient detail*

☐No

☐N/A

☐Unable to code

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

☒Yes *Explicit sequence used to teach manipulative use under control of teacher – modelling, guided practice, independent practice*

☐No

☐N/A

☐Unable to code

Overt measurement of the fidelity of implementation for the independent variable is highly desirable

☒Yes *Treatment fidelity evaluated during 25% of the sessions; observer measured whether teaching sequence was followed according to the procedure for each student. 100% adherence = close fidelity*

☐No

☐N/A

☐Unable to code

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

This rating was given as all criteria were fulfilled: the IV is described with precision, it is systematically manipulated and was carried out with high fidelity to the teaching sequence.

3. 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 *3 measurements at baseline for each DV for first child. As implementation of the intervention is staggered, there are 3+ baseline data points for the other children. All children scored 0 at baseline*

☐No

☐N/A

☐Unable to code

Baseline conditions are described with replicable precision

☒Yes *Baseline conditions described and achieved by administering same measures as for intervention phase*

☐No

☐N/A

☐Unable to code

Overall Rating of Evidence: ☒3 ☐2 ☐1 ☐0
This rating was given as all criteria were fulfilled: baseline conditions were described and sufficient data points were collected

4. Experimental Control/Internal Validity
The design provides at least three demonstrations of experimental effect at three different points in time
☒Yes *Staggered implementation of intervention for each child gives 3+ demonstrations of experimental effect. 3 time points: intervention, maintenance and long term.*
☐No
☐N/A
☐Unable to code

The design controls for common threats to internal validity (e.g. permits elimination of rival hypotheses)
☐Yes
☒No *There may have been bias is selection as 1 criteria was identification by the teacher as needing area and perimeter solving skills and this same teacher taught the intervention sequence*
☐N/A
☐Unable to code

The results document a pattern that demonstrates experimental control
☒Yes *There is little overlap between baseline and intervention and each participant acted as their own control*
☐No
☐N/A
☐Unable to code

Overall Rating of Evidence: ☐3 ☒2 ☐1 ☐0
This rating was given as there is sufficient demonstration of experimental effect and control. However, threats to internal validity were not considered or noted and there may have been extraneous variables impacting on the results.

6. Social validity
Experimental effects are replicated across participants, settings or materials to establish external validity
☒Yes *All three participants improved their area and perimeter problem solving performance*
☐No
☐N/A
☐Unable to code

The dependent variable is socially important

☒Yes *Being able to answer area and perimeter questions correctly after scoring 0 pre-intervention is likely to have a positive impact on the participants' self-concept in maths. Verbal feedback from the participants indicated they were positive about using the manipulatives and wanted to work with them each session.*
☐No
☐N/A
☐Unable to code

The magnitude of change in the dependent variable resulting from the intervention is socially important
☐Yes
☐No
☐N/A
☒Unable to code

Implementation of the independent variable is practical and cost effective
☐Yes
☐No
☐N/A
☒Unable to code *Information relating to cost is not given*

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
☒Unable to code

Overall Rating of Evidence: ☒3 ☐2 ☐1 ☐0
Although some aspects of social validity were not considered in the study, the key criteria for a high weighting is that the effects are replicated across either participants or setting (Horner et al., 2005). This was the case as all three participants improved their performance on area and perimeter problems.

Average WoE A across the 6 judgement areas:
Sum of individual quality ratings ÷ number of areas
16 ÷ 6 = 2.6
Overall WoE A = 2.6