Case Study 1: An Evidence-Based Practice Review Report

Theme: School Based Interventions for Learning

How effective are early intervention maths programmes in improving numeracy attainment in at risk pupils aged 4-7 years?

Summary

This systematic literature review aims to discover how effective early intervention maths programmes are in improving numeracy attainment in 'at risk' pupils. Previous research has suggested early intervention maths programmes are any structured programme implemented at school (in addition to usual maths instruction) by a trained professional (Gersten et al., 2007), delivered to those at a young age (age 4-7 years) (Clements & Sarama, 2011; Bryant et al., 2008) who may be 'at risk' of mathematical difficulties (Toll & Van Luit, 2012; Clements & Sarama, 2011). Seven studies met the inclusion criteria and were reviewed using the Weight of Evidence Framework (Gough, 2007) and the APA Task Force Coding Protocol by Kratochwill (2003). Early intervention maths programmes were effective in improving numeracy knowledge and attainment in 'at risk' pupils as indicated by medium-large effect sizes across the majority of studies, post intervention and follow up. They were most effective when they targeted the most 'at risk' pupils and taught basic number concepts. Recommendations for future research are discussed.

Introduction

Description of early intervention maths programmes

In this review, early intervention maths programmes are understood to be any structured programme implemented at school (in addition to usual maths instruction) by a trained professional, delivered to a small group of ‘at risk’ children to improve their mathematical knowledge and attainment. This was decided as, research has shown that ‘high risk’ children benefit most from intensive (at least 30minutes) instruction from a professional, in small groups of three to six children (Gersten et al., 2007). Early intervention in this review refers to interventions which target pupils at a young age (4-7 years) (Clements & Sarama, 2011; Bryant et al., 2008). Targeting those at a young age is important as disparities in mathematics achievement evident upon school entry increase as students advance through the school system; therefore the gap between high and low achievers becomes greater over time (Fogelman, 1983).Thus targeting those early on hopes to reduce this gap. Furthermore early interventions target those who have been identified as being ‘at risk’ of mathematics difficulties, either due to their underachievement in maths which acts as a precursor for later maths difficulties (Toll & Van Luit, 2012) or those from low socio-economic status (SES) backgrounds which may predispose them to having mathematics difficulties (Clements & Sarama, 2011).

Psychological theory

The standardised content outlined in the Primary National Strategy (2006) can lead to an assumption that children are developmentally ready to engage with formal number operations as they progress into Key Stage 1. However, much research suggests that children’s understanding of number concepts at this age progresses through developmental stages of cognitive restructuring rather than a gathering of knowledge, and this development does not occur at the same pace for all children (Steffe, Von Glasersfeld, Richards, & Cobb., 1983). Steffe et al., (1983) define five stages of number development (see Table 1).

Table 1 – Five stages of number development (Steffe et al., 1983)

Stage	***Definition***
1. Emergent Number stage	Development of counting sequence and 1:1 correspondence.
2. Perceptual Number stage	Addition of visible quantities.
3. Figurative Number stage	Children use number sequence logic and ordinal strategies to complete 'screened' addition but do not yet have sound understanding of mathematics operations.
4. Initial Number stage	Children have adult-like understanding of number and can comprehend symbols and operations.
5. Facile Number stage	Higher order number concepts are developing; children can focus on relationships between number and operations.

According to Steffe et al., (1983) children need to reach the 'Initial Number Stage' before they are able to successfully engage with the curriculum. Clements and Sarama (2007) support this suggesting that most children move from understanding core knowledge of numbers, to number relations to number operations (National Research Council, 2009). It is important to recognise that children progress through stages when learning maths and differences in progress are associated with previous experience, learning opportunities and intrinsic abilities.

Number sense is a term which has been used when researching maths development and refers to a well organised conceptual framework of number information that enables a person to understand numbers and number relationships to solve mathematical problems (Bobis, 1996). Good number sense is important because it lays the foundation for more advanced skills to develop. Deficient number sense has been found to be a core marker for severe and persistent learning difficulties in mathematics (Mazzocco, Feigenson & Halberda, 2011). Children should be provided with the necessary opportunities to develop through these stages, particularly those who are 'at risk' so they can acquire sufficient number sense to form a basis for subsequent maths learning.

To support transition through these stages research suggests that explicit and systematic instruction is critical to support students 'at risk' in mathematics (Baker, Gersten, & Lee, 2002). Explicit and systematic instruction is based on an approach in which instruction is teacher-directed. The learning environment is thus key in supporting children to develop through these stages. To provide the most effective instruction, research suggests that 'high risk' children benefit most from intensive instruction in small groups (Gersten et al., 2007). Effective instruction thus should be employed to bring about change and work to prevent the mathematical difficulties which some may face.

Rationale

In recent years, the education systems in many Western nations have put a greater focus on mathematics education. Mathematics difficulties have been increasingly recognised as problematic. The Cockcroft Committee (DES, 1982) emphasized mathematics education is needed for adult life and employment; is a valuable means of communication; and is an important tool for studying other subjects.

The importance of early intervention has gained increasing attention as early education experiences have an impact on later outcomes (Sylvia, 2009). Research has reported on the large individual differences which exist among children in their early numerical skills prior to starting formal education (Aunio, Hautamäki, Sajaniemi, & Van Luit, 2009). Furthermore, a meta-analysis by Ramey and Ramey (1998) found that interventions that began earlier in development afforded greater benefits. Therefore strengthening the numeracy skills of students in their early school years could be beneficial to students' mathematics learning in the long-term.

Early intervention is therefore clearly important, in particular for pupils 'at risk' of mathematical difficulties as there is a greater risk for them not developing mathematics automaticity that is required to become proficient in mathematics (Bryant et al., 2008). Pupils 'at risk' of mathematical difficulties are often identified after starting formal education due to their attainment in maths. However, it is also recognised that those from low SES backgrounds can also be 'at risk'. A disproportionate number of children from low-income families start school without the skills necessary for success in formal mathematics and are overrepresented in the population of students with diagnosed mathematics difficulties (Jordan & Levine, 2009). Although both low and high income groups of children may have informal experiences with quantitative situations, those from low SES backgrounds may have fewer opportunities to mathematize this tacit knowledge; that is, to reflect on and represent the situations (Clements, Sarama, Spitler, Lange, & Wolfe, 2011).

Despite this increasing recognition of the importance of early intervention maths programmes for those children ‘at risk’, few such intervention curricula are available, and evidence of their effectiveness is limited. This therefore forms the basis for this review as Gersten, Rolfhus, Clarke, Decker and Dimino (2015) suggests, a serious evaluation of a promising intervention curriculum in mathematics for students in the early grades is needed.

Relevance to Educational Psychology Practice

This area is of clear importance in Educational Psychology as in practice Educational Psychologists (EPs) can suggest evidence based interventions which have the potential to minimise the negative impact experienced by those with mathematics difficulties. Early intervention and prevention is a valuable aspect of EP practice (Farrell et al., 2006) and they can highlight the importance of early maths interventions, which will reduce difficulties in the long term. Furthermore EPs can communicate to schools the challenging nature of the curriculum (being highly organised, sequential and progressive). They can emphasize the importance of learning simpler elements successfully before moving on to others. Chinn and Ashcroft (1998) suggest, maths is a subject where one learns the parts; the parts build on each other to make a whole; knowing the whole enables one to reflect with more understanding on the parts, which in turn strengthens the whole. Teaching staff should be aware that without being secure in basic mathematical concepts children will find it difficult to progress. Moreover EPs can influence the support that schools provide to their most vulnerable ‘at risk’ pupils, to make a huge difference to the life chances of children and young people.

Review question

How effective are early intervention maths programmes in improving numeracy attainment in at risk pupils aged 4-7 years?

Critical review of the evidence

Literature search

Initial searches were conducted during December, 2015 using electronic databases ERIC, PsycINFO, and Web of Science to identify all studies relevant to the review question. Table 2 shows the search terms used in order to identify relevant studies.

Table 2 – search terms used to locate studies on databases.

Search Terms	ERIC	PsycINFO	Web of Science
Numeracy intervention AND group	87	8	87
Mathematics intervention AND small group AND at risk	20	1	17
Mathematical intervention programmes	19	0	2

Inclusion and exclusion criteria

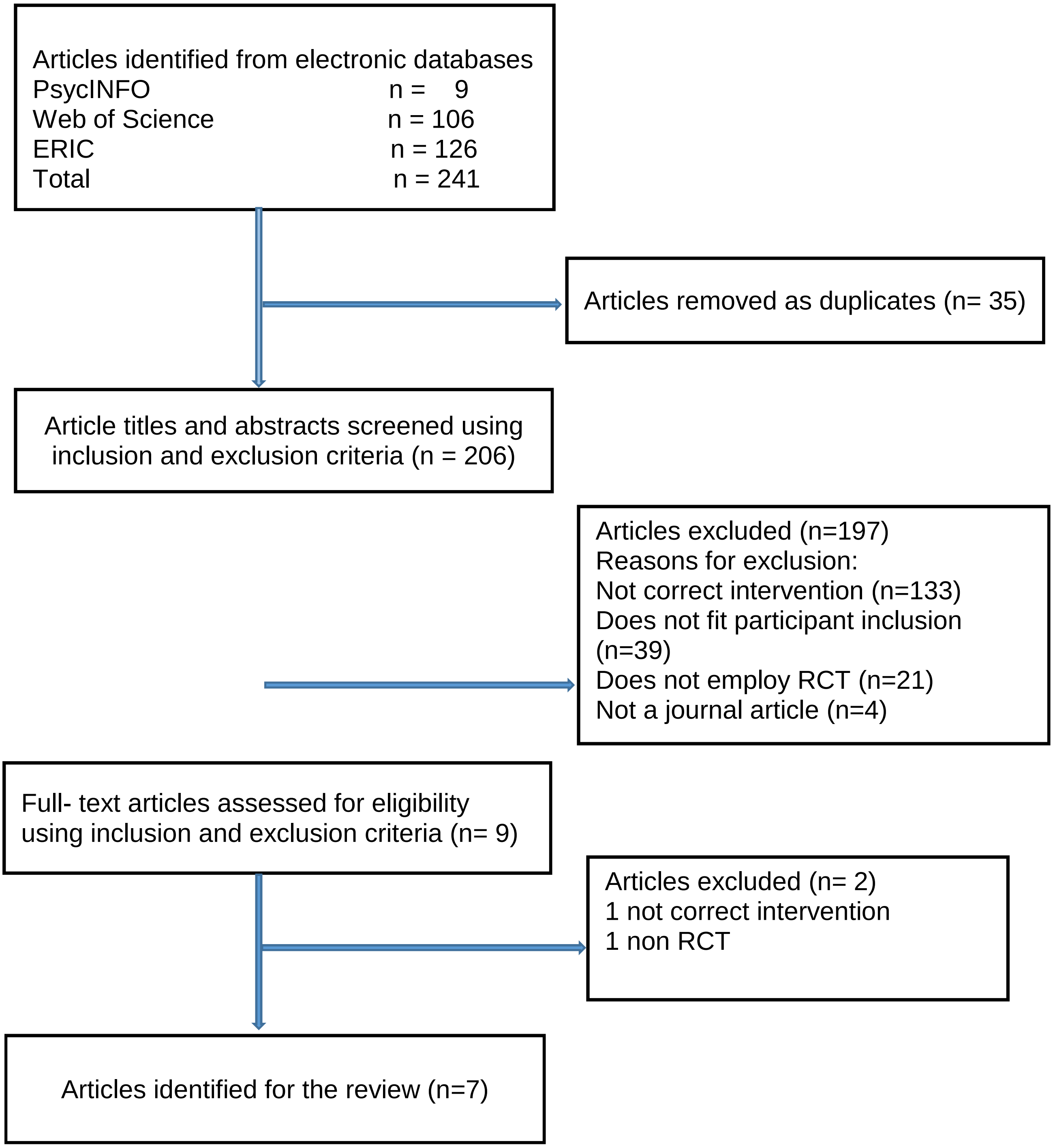
Studies retrieved were included in the review if they met the criteria detailed in Table 3. As shown in Figure 1, 241 studies were gathered through the electronic databases using the search terms in Table 2: 126 (ERIC), 9 (PsycINFO) and 106 (Web of Science). Out of the 241 studies obtained, 35 were removed as duplicates and subsequently excluded. The titles and abstracts of the remaining 206 studies were screened adhering to the inclusion criteria specified in Table 3. Those not meeting the inclusion criteria (n=197) were excluded on this basis. This resulted in 9 articles which required a full text inspection to see if they met the inclusion criteria. Of these, 2 did not meet the specified criteria and were excluded (see Appendix A for reasons of exclusion). As shown in Appendix B a total of 7 articles were suitable for inclusion for in-depth review.

Table 3 – Inclusion and exclusion criteria.

	Inclusion criteria	*Exclusion criteria*	*Rationale*
1. Type of publication	a) Must be peer reviewed.	a) Not peer reviewed	Peer reviewers assess the quality of a study and therefore the studies are likely to meet the required standards.
2. Language and setting	a) Must be written in English. No restrictions on country in which the study has taken place.	a) Study is not written in English.	Reviewer does not have resources to access other languages.
3. Type of design	a) Must be a RCT.	b) The study is not an RCT.	RCT's are considered the gold standard as participants are randomly allocated and bias is reduced.
4. Intervention	a) Must target maths achievement. b) Must be an early intervention (intervention takes place early on in the child's life age 4-7 years). c) Must target 'at risk' pupils	a) Does not target maths achievement. b) Not an early intervention (takes place when the child is 8 and older).	Early interventions lack research. There is evidence that interventions which target those 'at risk' can reduce difficulties later in life. Small group work has been found to be beneficial as a targeted intervention when delivered by a trained

	Inclusion criteria	*Exclusion criteria*	*Rationale*
	either due to academic performance or low SES. d) Must be a small group intervention (2-5 pupils). e) Must be explicit teaching delivered by trained adult.	c) Does not target 'at risk' pupils. d) Individual intervention or curriculum wide intervention. e) Not delivered by trained adult i.e. delivered by computerised programme or playing games.	professional.
5. Participants	a) Must be attending school. b) Aged 4-7 years. c) Identified as being 'at risk' of mathematical difficulties due to academic performance or low SES. d) Normal developing children.	a) Not attending school b) Not in the 4-7 year age range. c) Not 'at risk' of mathematical difficulty. d) Children with additional needs or SEN.	Review aims to identify whether early intervention maths programmes are effective in improving maths attainment in 'at risk' pupils.

Figure 1- Flowchart documenting the application of inclusion and exclusion criteria



Critical appraisal for quality and relevance

The seven papers were summarised to provide an overview and to examine their relevance to the review question (See Appendix C). The quality and relevance of each study was appraised using the Weight of Evidence (WoE) framework (Gough, 2007). This framework considers three dimensions for weighting studies: WoE A, B and C and then considers all these ratings together to rate the overall WoE (WoE D) of each identified study. WoE A examines methodological quality in terms of the quality of execution of the study in relation to quality standards for studies of that type. WoE A in this review was judged using an adapted version of the APA Task Force Coding Protocol by Kratochwill (2003). The protocol was adapted to ensure it was appropriate to the review question and was then used to judge methodological quality (See Appendix D). Each study was examined using this protocol in order to evaluate their methodological quality based on the same criteria. WoE B examines the methodological relevance of the studies in relation to the review question and WoE C appraises the topic relevance of the studies in relation to the review question. The outcomes of WoE A, B and C are averaged to provide WoE D (see Table 4). Information on how each study was appraised is detailed in Appendix E.

Table 4- Overall Weight of Evidence

Studies	WoE A: Quality of methodology	WoE B: Relevance of methodology	WoE C: Relevance of evidence	WoE D: Overall Weight of Evidence
Dyson et al., (2015)	High 2.6	High 3	Medium 2	High 2.5
Gersten et al., (2015)	Medium 2.4	Medium 2	Medium 2	Medium 2.1
Toll and Van Luit (2012)	Medium 1.6	Medium 2	Medium 2	Medium 1.9

Studies	WoE A: Quality of methodology	WoE B: Relevance of methodology	WoE C: Relevance of evidence	WoE D: Overall Weight of Evidence
Jordan et al., (2012)	High 3	High 3	Medium 2	High 2.7
Bryant et al., (2011)	Medium 1.8	Medium 2	Medium 2	Medium 1.9
Dyson et al., (2011)	Medium 2.4	High 3	Low 1	Medium 2.1
Aunio et al., (2005)	Medium 1.8	Medium 2	Medium 2	Medium 1.9

Participants

The number of participants in the studies ranged from 45-994. In line with the inclusion criteria, all participants were primary school children attending mainstream settings: aged from 4-7 years. The participants were all typically-developing and were identified as being 'at risk' based on coming from low SES background (Jordan et al., 2012; Dyson et al., 2011) or through a maths attainment measure (Dyson et al., 2015; Gersten et al., 2015; Toll & Van Luit, 2012; Bryant et al., 2011 and Aunio et al., 2005). Participants were recruited from the United States (n=5 studies), The Netherlands (n=1 study) and Finland (n=1 study). All except two studies recruited participants from low SES backgrounds (Aunio et al., 2005 and Toll & Van Luit, 2012). While low SES may predispose some children to mathematical difficulties, it may not be the case for all. Therefore those studies which identified children from low SES backgrounds and then undertook a maths attainment measure were rated more highly for WoE C than those interventions that targeted all children from low SES backgrounds.

Whilst those studies which identified children from low SES backgrounds and then undertook a maths attainment measure were rated more highly for WoE C, the appropriateness of the ‘cut-off’ criteria selected by these studies to identify those ‘at risk’ must be considered. The studies which employed this method to identify those at risk (Dyson et al., 2015; Gersten et al., 2015; Toll & Van Luit, 2012 and Aunio et al., 2005) all differed in their ‘cut-off’ criteria. For example, Dyson et al., (2015) identified those ‘at risk’ if they achieved below the 25th percentile, whereas Toll and Van Luit (2012) identified ‘at risk’ pupils if their norms were below the 50th percentile. These differences in criteria may lead to different levels of need being targeted. While this appears to be the most appropriate method to select potential ‘at risk’ pupils, difficulties with defining this population remain. It is unknown whether these pupils who are identified as underachieving are ‘at risk’ or whether they already have difficulties with maths. The literature does however suggest that targeting these difficulties early on at a young age will have a strong positive effect on the child’s academic life and therefore reduce risk for difficulties at a later age (Clements & Sarama, 2008).

Samples were attained by convenience sampling methods, wherein schools could voluntarily take part in the studies. After schools agreed to take part, parents then provided consent. In psychological research this is a common technique to acquire large numbers of participants, however a limitation of this non-random sampling is that not every member of the target population has an equal opportunity of being selected (Barker, Pistrang & Elliot, 2002). This may lead to bias within the research, particularly when schools voluntarily have the option to participate. Mathematics may have been an area of the curriculum that the school were working to improve and they may have been more invested in maths attainment so provided increasing support outside of the intervention. Furthermore, parents may have particular characteristics (for example, motivation for their child to improve) and subsequently may have been supporting them more at home. This should be considered in relation to generalizability of results to the target population.

Intervention

A range of early maths intervention programmes were used to support children identified 'at risk' for mathematics difficulties. The majority of the studies focused on targeting children's early number sense (Aunio et al., 2005; Dyson et al., 2011; Jordan et al., 2011; Gersten et al., 2015 and Dyson et al., 2015). Whilst interventions targeted similar concepts such as number recognition, number relations and problem solving they were often taught differently. This makes it difficult to compare efficacy, however all studies did describe how the intervention was implemented. In all studies a trained professional provided the intervention to a small group of children and it was always in addition to their maths lesson as usual. Those studies which provided implementers with ongoing supervision (Gersten et al., 2015; Jordan et al., 2012 and Bryant et al., 2011) and ensured manualization (Dyson et al., 2015; Toll & Van Luit, 2012 and Dyson et al., 2011) were rated more highly for WoE A as it increased fidelity of implementation.

The interventions in all the studies lasted between 25-40 minutes for each session, and sessions received ranged from 2-4 per week. Duration of the intervention also differed, four interventions lasted for 8 weeks (Dyson et al., 2015; Toll & Van Luit, 2012; Jordan et al., 2012; and Dyson et al., 2011) whilst the others varied from 17 weeks (Gersten et al., 2015) 19 weeks (Bryant et al., 2011) and 9 months (Aunio et al., 2005). Therefore it is important to consider the length of the intervention may have had an impact on the subsequent outcomes.

Design

All studies included in this review employed randomised controlled trials (RCTs). The use of a control group increases the internal validity of findings and conclusions that can be drawn, since it makes the distinction between changes that occur as a result of the intervention and that of endogenous change. In all studies they employed an 'active' control group who received the maths curriculum as usual and any normal classroom support. This allowed researchers to recognise whether those identified as 'at risk' would benefit more from targeted small group maths support in addition to normal maths instruction, compared to normal maths instruction alone.

Randomisation methods used in the studies reviewed varied: the majority randomised students to conditions but two randomised at the school level to conditions (Gersten et al., 2015 and Toll & Van Luit, 2012).This allowed each participant or school to have an equal chance of being in each condition. This review is interested in reducing maths difficulties in 'at risk' pupils, therefore the extent to which effects are maintained is important, and studies which examined outcomes at follow up therefore received a higher WoE. Only four studies conducted a follow up (Dyson et al., 2015; Jordan et al., 2012; Dyson et al., 2011; and Aunio et al., 2005). Those studies which did not include follow up (Gersten et al., 2015; Toll & Van Luit, 2012 and Bryant et al., 2011) only received a 'medium' rating for WoE B and C as they only examined the immediate effects.

Measures

All outcome measures assessed early mathematics ability, however measures employed differed across studies. The Number Sense Brief (NSB; Jordan, 2010) was used most frequently as it was employed in three studies (Dyson et al., 2015; Jordan et al., 2012; and Dyson et al., 2011). Other measures included, Early Numeracy Test (ENT; Van Luit et al., 2003) used by Aunio et al., (2005) and Early Numeracy Test-Revised (ENT-R; Van Luit et al., 2009) employed by Toll and Van Luit (2012). Test of Early Mathematics Ability III (TEMA-3; Ginsburg et al., 2003) was used by Gersten et al., (2015) and Bryant et al., (2011) used the Texas Early Mathematics Inventories Outcome (TEMI-O; University of Texas System and Texas Education Agency, 2007). To assess general mathematics achievement some studies employed the Woodcock-Johnson III Tests of Achievement (WJ-III) Standard Test Calculation and Applied problems subtests (Woodcock, McGrew, & Mather, 2007). More detail regarding these measures is presented in Table 5. All measures used had high internal consistency and this is reflected in WoE A.

The majority of studies also employed measures to assess a range of background variables including attention, reading achievement, working memory and general thinking abilities. These measures have not been included in the analysis as the key focus of this review was related to improving maths attainment. This review has therefore focused the analysis to include the main outcome measures in the studies which focused on early maths ability and achievement.

The outcome measures used assessed early maths ability and achievement pre-post and pre- follow up intervention compared to the control group. As the measures were not from multiple sources (for example, teacher ratings, parent ratings) it is possible that effects found using the outcome measures did not generalise to the children's performance in their maths lessons or their homework. This is therefore reflected in WoE A. It is important that the effects demonstrated are generalizable to other domains as the interventions are targeting 'at risk' pupils who will need to be able to cope with curriculum demands.

Table 5 – Information about the measures used in the studies included in this review.

Measure	Details of measure	Subtests	Studies employing measure	Internal consistency
The Number Sense Brief (Jordan, Glutting, Ramineni, & Watkins, 2010)	Composed of a range of questions to assess a child's number knowledge. Every item is presented orally to the child, and the child gives an oral response. The measure is untimed and takes approximately 20 minutes to complete.	Counting, number recognition, number comparisons, nonverbal calculation, story problems, and number combinations.	Dyson et al., (2015) Jordan et al., (2012) Dyson et al., (2011)	0.90
Woodcock Johnson III tests of achievement Form C Brief Battery- Applied Problems and Calculation subtests (Woodcock, McGrew, & Mather, 2007).	The Applied Problems subtest requires children to solve orally presented problems through the application of quantitative reasoning. It also assesses money and measurement concepts. The Calculation subtest assesses computation in a written format with conventional symbols.	Applied problems and Calculation.	Dyson et al., (2015) Jordan et al., (2012) Dyson et al., (2011)	0.90
The Early Numeracy Test (ENT) and The Early Numeracy Test-Revised (Van Luit & Van de Rijt, 2009)	Assesses math prerequisites, counting, and estimation skills.	Comparison, classification, correspondence, seriation, using numerals, synchronized and shortened counting, resultative counting, general understanding of numbers, and estimation.	Toll and Van Luit (2012) Aunio et al., (2005)	0.82

Measure	Details of measure	Subtests	Studies employing measure	Internal consistency
Test of Early Mathematics Ability III (Ginsburg & Baroody, 2003)	Measures both formal and informal mathematics knowledge and skills. The content assessed by the TEMA-3 is designed to be consistent with typical Grade- level curricula taught in schools in the area of number.	Numbering, number comparisons, calculation, concepts, numeral literacy, number facts, and calculation.	Gersten et al., (2015)	0.95
Texas Early Mathematics Inventories Outcome (University of Texas System and Texas Education Agency, 2007).	A group administered problem-solving and whole-number computation measure. The TEMI-O was considered to be a distal outcome measure because it assesses all of the state's standards for first-grade instruction.	Composed of two subtests: Mathematics Problem Solving (MPS) and Mathematics Computation (MC).	Bryant et al., (2011)	0.86

Findings

Effect size data is summarised in Table 6 and a more in depth analysis can be found in Appendix F. The studies all found some significant positive effects of early intervention maths programmes targeting those 'at risk' of mathematics difficulties when compared to the 'active' control groups.

Table 6 – Summary of effect size data from studies included in the review.

Study	Outcome measure	Effect Sizes Pre-post	 Pre- follow up	Overall Weight of Evidence
Dyson et al., (2015)	Number sense (NSB)			High
	Number list	g = .32 (small)	g = .26 (small)	
	Number fact	g = .82 (large)	g = .56 (medium)	
	Maths achievement (WJ-III)			
	Number list	g= .58 (medium)	g= .12 (small)	
	Number fact	g = .60 (medium)	g =.49 (small)	
Jordan et al., (2012)	Number Sense (NSB)	d = 1.80 (large)	d = .63 (medium)	High
	Maths achievement (WJ-III)	d = .91 (large)	d = .83 (large)	
Dyson et al., (2011)	Number Sense (NSB)	R^2 = .30 (large)	R^2 = .29 (large)	Medium
	Maths achievement (WJ-III)	R^2 = .16 (medium)		
Toll and Van Luit (2012)	Early numeracy skills (ENT-R)	η^2 = .40 (large) 25th-50th percentile η^2 = .22 (large)		Medium
Aunio et al., (2005)	Early numeracy skills (ENT)	η^2 = .13		Medium

Study	Outcome measure	Effect Sizes Pre-post	Pre- follow up	Overall Weight of Evidence
		(Medium) $\eta^2 = .22$ (Large)		
Gersten et al., (2015)	Mathematics ability (TEMA-3)	$g = .34$ (small)		Medium
Bryant et al., (2011)	Early mathematics ability (TEMI- O)	$g = .21$ (small)		Medium

Note. Cohen's *d* and Hedges's *g* effect sizes are considered small at 0.2, medium at 0.5 and large at 0.8 (Cohen, 1988). Partial eta-squared (η^2) critical values are 0.01 for a small effect, 0.06 for a medium effect and 0.14 for a large effect (Cohen, 1988). R^2 represents multiple correlation coefficients squared. Values equal to .02 are considered small, those at .13 are moderate and those at .25 are considered large effects (Cohen, 1988).

While all studies demonstrated significant positive effects of intervention, the size of these effects ranged considerably from large, to medium and small. Differences in the intervention implemented to support those 'at risk' and different outcome measures used to assess improvements may have impacted the effect sizes found. Three studies employed the NSB and therefore it is possible to compare the effect of these interventions more accurately. All found large effects post intervention, except Dyson et al., (2015) in which they compared two similar interventions which differed in the last 5 minutes of each session. The number fact intervention was more successful in improving children's number sense as opposed to the number list. A particular strength of these studies is their inclusion of follow up measures. Positive effect sizes did remain at follow up, however were reduced to medium in both Dyson et al., (2015) and Jordan et al., (2012) but were maintained in the Dyson et al.,

(2011) study. Whilst this maintenance is present, effects must be considered with the weightings of the studies. Dyson et al., (2015) and Jordan et al., (2012) both received high overall WoE whereas Dyson et al., (2011) only received a medium WoE. The fact that Dyson et al., (2015) and Jordan et al., (2012) found large effects at post and medium effects at 8 week follow up is positive. This suggests a short intervention (over a period of 8 weeks) can support children to improve their number sense and effects are maintained. Whilst Dyson et al., (2011) found similar effects, caution must be taken when interpreting the results. This intervention targeted all children from low SES and whilst this may predispose some children, not all may be 'at risk' of mathematical difficulties.

These studies also employed the WJ-III subtests and found effects ranged from medium to large. In Dyson et al., (2015) and Dyson et al., (2011) these effects significantly reduced at follow up. Jordan et al., (2012) however found large effects which were maintained at 8 week follow up. As a highly weighted study these findings are impressive and suggest children were able to internalize what they had learnt during the intervention and improve their maths attainment.

Only one other study included follow up data (Aunio et al., 2005) which took place 6 months post intervention. However, whilst medium and large effects were found for relational and counting subtests, these effects had faded at 6 months follow up. This suggests that while immediate effects are present, after a period of 6 months with no additional intervention, effects were no longer present. This is surprising considering the intervention in this study was in place for a period of 9 months. However, the study was underpowered and there was only a small number of participants identified as being ‘at risk’ (n=12) and as a result received a medium overall WoE. A potential for this may be that the intervention did not target children who needed to improve in their maths attainment. If they were already high achievers in maths they are unlikely to improve significantly. While this intervention seems appropriate for short term effects, the findings suggest it may not be as beneficial in the long term.

The remaining studies, did not include follow up measures and subsequently received a medium overall WoE. While immediate effects of the interventions were captured, for this review it was deemed important to collect follow up data to ensure the ‘at risk’ children would no longer be ‘at risk’. Toll and Van Luit (2012) found large effects at immediate post- test (that children had obtained higher adjusted outcome numeracy scores). Effects however were only found to be significant for those children who had a pre-test numeracy score falling between the 25th-50th percentiles, however for those lower achievers (scoring below 25^{th} percentile) effects were not found to be significant at post-test. This suggests children need to have acquired the basics before they can benefit from additional targeted instruction and it must be targeted at the correct level or the intervention will lack effectiveness.

Both Gersten et al., (2015) and Bryant et al., (2011) only found small effects post intervention and while they only received a medium overall WoE it is important to consider these findings. Both these studies employed broad outcome measures which are consistent with the state’s curriculum. As this review is exploring effectiveness of early maths interventions, for children ‘at risk’, the overarching aim is for them to be able to keep up with the curriculum demands. These findings suggest that interventions may not be targeted appropriately to ensure that children can meet the demands placed on them. It may be that interventions which teach the basics and target number sense, are more effective to ensure children can then progress through the curriculum.

The definition of 'at risk' pupils also differed across studies and this must be considered when interpreting findings. There was great variance in the mathematical ability of pupils included across the studies and this should not be overlooked. For example, one study identified those 'at risk' whose norms were below 50^{th} percentile on a maths attainment measure (Toll & Van Luit, 2012) whereas Gersten et al., (2015) identified those as 'at risk' if they scored below the 35^{th} percentile and Dyson et al., (2015) identified pupils who achieved below the 25^{th} percentile. Dyson et al., (2011) further differed as this study identified those 'at risk' if they came from a low SES background. The wide variance in 'at risk' definitions means that findings must be interpreted with caution as interventions may have potentially targeted different abilities and thus it makes comparison of these studies difficult.

Conclusions

This systematic literature review aimed to discover the effectiveness of early intervention maths programmes to target maths difficulties in 'at risk' pupils. Early intervention maths programmes that focused on early number sense and targeted basic concepts were found to support those 'at risk' with large to medium effects. Mixed effects however, are found at follow up. In the majority of cases whilst effects of the intervention declined after implementation, they were still significant suggesting that children had internalised knowledge. Findings suggest that interventions need to be targeted appropriately to the most ‘at risk’ pupils to ensure they have acquired a basic knowledge of mathematical concepts before progressing to more complex ideas. This review suggests that maths interventions provided by a trained professional, to a small group of ‘at risk’ children leads to improved maths knowledge and attainment.

Limitations

A key limitation is that not all studies included in the review examined the maintenance of the effects of the intervention. It is important to explore if effects are maintained at follow up, particularly for this population. Furthermore, in all studies included in the review, interventions were carried out by external, trained professionals. It would be beneficial to explore if effects would also be found if interventions were delivered by school staff. It is unknown whether effects found on outcome measures also transferred to other settings. Triangulation of data such as information from teachers and parents would be useful to understand if effects were generalised. The differing definitions used by each study to define ‘at risk’ is also a key limitation especially when comparing studies as interventions were potentially targeting children with different levels of maths ability.

Recommendations

In light of the limitations, whilst early intervention maths programmes for children ‘at risk’ show promising and beneficial immediate effects and some maintained effects at follow up, more research is necessary to examine these effects in the long term. More research which is able to demonstrate effects after a significant follow up period would be able to inform EP practice further. Longitudinal studies which examine the effect of early interventions over time would be useful to examine long term outcomes. This is necessary as when EPs advise schools on supporting their most vulnerable pupils they would be able to demonstrate evidence based research that can make a difference to these pupils' maths achievement throughout their school life and also their success in adulthood. Moreover, research which examines the effectiveness of school staff implementing early maths interventions is essential. Ensuring school staff take ownership of the intervention is important so that EPs can enable schools to build their own capacity. It is also important that in future research 'at risk' populations are defined with clarity to ensure interventions are targeted appropriately to meet the needs of the most vulnerable.

References

Aunio, P., Hautamäki, J., Sajaniemi, N., & Van Luit, J. E. H. (2009). Early numeracy in low-performing young children. *British Educational Research Journal*, 35(1), 25-46.

Baker, S., Gersten, R., & Lee, D.S. (2002). A synthesis of empirical research on teaching mathematics to low-achieving students. *Elementary School Journal*, 103, 51–73.

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

Bobis, J. (1996). Visualisation and the development of number sense with kindergarten children. In Mulligan, J. & Mitchelmore, M. (Eds.) *Children's Number Learning* : A Research Monograph of the Mathematics Education Group of Australasia and the Australian Association of Mathematics Teachers. Adelaide: AAMT

Bryant, D. P., Bryant, B. R., Gersten, R., Scammacca, N., Funk, C., Winter, A., Shih, M., & Pool, C. (2008). The effects of Tier 2 intervention on first-grade mathematics performance of first-grade students who are at risk for mathematics difficulties. *Learning Disability* Quarterly, 31, 47–63.

Chinn, S.J. and Ashcroft, J.R. (1998). Mathematics for Dyslexics: A Teaching Handbook. London: Whurr Publishing.

Clements, D. H., & Sarama, J. (2007). Effects of a preschool mathematics curriculum: Summative research on the Building Blocks project. *Journal for Research in Mathematics Education,* 38, 136–163.

Clements, D. H., & Sarama, J. (2011). Early childhood mathematics intervention. *Science*, 333, 968–970.

Clements, D. H., Sarama, J., Spitler, M. E., Lange, A. A., & Wolfe, C. B. (2011). Mathematics learned by young children in an intervention based on learning trajectories: A large-scale cluster randomized trial. *Journal for Research in Mathematics Education,* 42(2), 127–166.

Cohen, J. (1988). Statistical power analysis for the behavioural sciences. *Statistical Power Analysis for the Behavioural Sciences*.

Department for Education and Science (1982) *Mathematics counts* (Report of the Cockcroft Committee of Enquiry into the teaching of Mathematics). London: HMSO.

Department for Education and Skills (DfES). (2006). Primary National Strategy: Primary Framework for literacy and mathematics. London: DfES.

Evans, D. (2003). Hierarchy of evidence: A framework for ranking evidence evaluating healthcare interventions. *Journal of Clinical Nursing*, 12, 77–84.

Farrell, P., Woods, K., Lewis, S., Rooney, S., Squires, G., & Connor, M. O. (2006). A Review of the Functions and Contribution of Educational Psychologists in England and Wales in light of “Every Child Matters: Change for Children.”

Fogelman, K. (Ed.). (1983). Growing up in Great Britain: Papers from the national child development study. London: Macmillan.

Gersten, R., Baker, S. K., Shanahan, T., Linan-Thompson, S., Collins, P., & Scarcella, R. (2007). *Effective literacy and English language instruction for English learners in the elementary grades: A practice guide* (NCEE 2007-

4011). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences,U.S. Department of Education.

Gersten, R., Rolfhus, E., Clarke, B., Decker, L. E., & Dimino, J. (2015). Intervention for First Graders with Limited Number Knowledge: Large-Scale Replication of a Randomized Controlled Trial, *American Educational Research Journal*, *52*(3), 516–546.

Ginsburg, H., & Baroody, A. (2003). Test of Early Mathematics Ability–third edition. Austin, TX: Pro-Ed.

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.

Jordan, N. C., & Levine, S. C. (2009). Socioeconomic variation, number competence, and mathematics learning difficulties in young children. *Developmental Disabilities Research Reviews*, 15, 60–68.

Jordan, N. C., Glutting, J., Ramineni, C., & Watkins, M. W. (2010). Validating a number sense screening tool for use in kindergarten and first grade: Prediction of mathematics proficiency in third grade. *School Psychology Review,* 39(2), 181–195.

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

Mazzocco, M. M. M., Feigenson, L., & Halberda, J. (2011). Impaired acuity of the approximate number system underlies mathematical learning disability. *Child Development,* 82, 1224–1237.

National Research Council. (2009). Mathematics learning in early childhood: Paths toward excellence and equity. Washington, DC: National Academies Press.

Ramey, C. T., & Ramey, S. L. (1998). Early intervention and early experience. *American Psychologist*, 53(2), 109-120.

Steffe, L. P., Von Glasersfeld, E., Richards, J., & Cobb, P. (1983). Children's counting types: Philosophy, theory, and application. New York: Praeger.

Sylvia, K. (2009). Early childhood matters: Evidence from the effective preschool and primary education project. London: Taylor and Francis.

University of Texas System & Texas Education Agency. (2007). *Texas early mathematics inventories-outcome.* Austin, TX: Authors.

Van Luit, J. E. H., Van de Rijt, B. A. M. & Aunio, P. (2003) *Early numeracy test, Finnish edition* [*Lukukäsitetesti*] Helsinki: Psykologien kustannus.

Van Luit, J. E. H., & Van de Rijt, B. A. M. (2009). *Utrechtse getalbegrip toets — revised*. [Early Numeracy Test–Revised]. Doetinchem, Netherlands: Graviant.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2007). *Woodcock-Johnson tests of achievement.* Itasca, IL: Riverside Publishing.

Appendix A - List of Excluded Studies

The studies listed below came up in the searches conducted in ERIC, PsycINFO and Web of Science. They were excluded when the reviewer read past the title and abstract to the main content of the study. Studies that came up in searches but did not match the inclusion criteria due to information in their title or abstract are not included in this list.

Studies	Reason for exclusion
Somerville, R., Ayre, K., Tunbridge, D., Cole, K., Stollery, R., Somerville, R., Stollery, R. (2015). Firm foundations : the effectiveness of an educational psychologist developed intervention targeting early numeracy skills, *Educational Psychology in Practice,* 31:3, 265-278.	Does not meet inclusion criteria 3a – the study is not a RCT.
Young-Loveridge, J. M. (2004). Effects on early numeracy of a program using number books and games. *Early Childhood Research Quarterly*, 19(1), 82–98.	Does not meet inclusion criteria 4d- the intervention focuses on playing games.

Appendix B – Included studies

Included studies

Aunio, P., Hautamäki, J., & Luit, J. E. H. Van. (2005). Mathematical thinking intervention programmes for preschool children with normal and low number sense*, European Journal of Special Needs Education,* 20:2, 131-146.

Bryant D.P., Bryant B. R., Roberts, G., Vaughn, S., Hughes Pfannenstiel, K., Porterfield, J., Gersten, R., (2011). Early Numeracy Intervention Program for First-Grade Students with Mathematics Difficulties, *Exceptional Children*, *78*(1), 7–23.

Dyson, N. I., Jordan, N. C., & Glutting, J. (2011). A Number Sense Intervention for Low- Income Kindergartners at Risk for Mathematics Difficulties, *Journal of Learning Disabilities,* 46(2), 166-181.

Dyson, N., Jordan, N. C., Beliakoff, A., & Hassinger-Das, B. (2015). A kindergarten number-sense intervention with contrasting practice conditions for low-achieving children. *Journal for Research in Mathematics Education*, 46(3), 331–370.

Gersten, R., Rolfhus, E., Clarke, B., Decker, L. E., & Dimino, J. (2015). Intervention for First Graders with Limited Number Knowledge: Large-Scale Replication of a Randomized Controlled Trial, *American Educational Research Journal*, *52*(3), 516–546.

Jordan, N. C., Glutting, J., Dyson, N., Hassinger-das, B., & Irwin, C. (2012). Building Kindergartners’ Number Sense : A Randomized Controlled Study, *Journal of Educational Psychology*, 104(3), 647–660.

Toll, S.W.M., & Van Luit, J.E.H. (2012). Early Numeracy for Low-Performing Kindergartners, *Journal of Early Intervention*, 34(4), 243-264.

Appendix C – Mapping the field.

	Author	Sample	Study Design	Age, gender, presenting difficulty	Country	Intervention
1.	Dyson et al., 2015	44-number fact practice 40- number list practice 42- control	RCT	Kindergarten pupils (5-6 years) Male=95 Female=93 Those identified as 'at risk' were done so via the number sense screener and those scoring below the mean raw score (<21) were identified as 'at risk'.	United States	Number sense intervention- (4 pupils per group) carried out over 24, 30min sessions 3 times a week in addition to maths instruction as usual. Both conditions received the same lesson, only the last 5 min differed. Number fact- encouraged quick answers in response to questions. Number list- played a number game.
2.	Gersten et al., 2015	615- intervention 379- control condition	RCT	First grade students (6-7 year olds). Female= 47% intervention group. A composite score for each student was computed based on average z-scores across six subtests and implemented a cut off score at the 35th percentile of the sample to identify students	United States	Number rockets- Pupils (2 or 3 per group) participated 3 or more times per week for approx. 17 weeks in addition to normal maths lesson. Each session lasted 40mins (30mins instruction 10 min fact practice).

Author	Sample	Study Design	Age, gender, presenting difficulty	Country	Intervention	Implementer	Follow up
			'at risk' for mathematics difficulties.				
3. Toll and Van Luit, 2012	52- intervention 46- control 94- typically achieving	RCT	Kindergarten pupils in their second year (mean age 5.23 years) Male=100 Female= 96 Children with a pre-test numeracy score falling below the 50th percentile were identified as 'at risk'.	The Netherlands	The Road To Mathematics (TRTM) program offered in 16 approximately 30 minute sessions twice a week. 3 or 4 children participated in each group in addition to normal maths lesson.	4 trained supervisors who had at least Bachelor's degree in education, special education or child psychology and all received training to deliver the intervention.	No follow up
4. Jordan et al., 2012	42- intervention 42- language group 44- control	RCT	Kindergarten pupils (5-6 years) Male=65 Female=63 Pupils from low income communities were identified as 'at risk' for underachievement in maths.	United States	Number sense intervention- 30 minute lessons in groups of 4, three times a week in addition to normal maths lesson.	6 university graduates and 5 undergraduates who were trained to deliver number sense intervention.	8 week follow up
5. Bryant et al., 2011	139- intervention 65- control	RCT	First grade pupils (6-7 years) Male=43.9% Female=56.1% Pupils identified as 'at risk' based on their score (lowest 35%) on Texas Early Mathematics	United States	Early numeracy preventative Tier 2 intervention- occurred four times a week for 25 minute sessions over 19 weeks. 3-5 pupils per group were tutored.	Trained tutor from the research team delivered the intervention daily.	No follow up

Author	Sample	Study Design	Age, gender, presenting difficulty	Country	Intervention	Implementer	Follow up
			Inventories Progress Monitoring Measures (TEMI-PM).				
6. Dyson et al., 2011	56- intervention 65- control	RCT	Kindergarten pupils (5-6 years) Male=69 Female=52 Pupils were identified as at risk for mathematics underachievement as they were from low income families.	United States	Number sense intervention- groups of 4 pupils met for 3 30minute sessions a week for a period of 8 weeks in addition to normal maths instruction.	6 university students studying education who were trained in carrying out the intervention.	6 week follow up.
7. Aunio et al., 2005	22- intervention 23- control	RCT	Pre-schoolers (4-6 years). Male=27 Female=18 Children identified as having low number sense by applying the criterion of a minus- 1sd score below the total score mean in the *Early numeracy test. (*12 low performers identified).	Finland	‘Let’s think!’ and ‘Young children with special educational needs count, too!’- two sessions a week for 30 minutes, for a period of 9 months	Author of the paper and two early childhood special educators.	6 month follow up

Appendix D – An example coding protocol

Coding protocols and items removed from APA Task Force Coding Protocol by **Kratochwill (2003**)

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**: 17/1/16 ____________________

Full Study Reference in APA format: Dyson, N., Jordan, N. C., Beliakoff, A., & Hassinger-Das, B. (2015). A kindergarten number-sense intervention with contrasting practice conditions for low-achieving children. *Journal for Research in Mathematics Education*, *46*(3), 331–370.

Intervention Name (description from study): Number sense intervention followed by either a number-fact practice session or followed by a number-list practice session.

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)
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)
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)
A3.4 ☐ High (strong inference)
A3.5 Very high (explicit)
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: ANCOVA
_level: 0.5

ES: $f_2 = .15$
N required: 126

B4. Total size of sample (start of the study): 132
N

B5. Intervention group sample size: 40 Number list
44 Number-fact

B6. Control group sample size: 42
N

C. Type of Program (select one)

C1. ☐ Universal prevention
☐ Selective prevention
☐ Targeted prevention
☐ Intervention/Treatment
☐ 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

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)
B2.4 High (strong)
B2.5☐ Very high
B2.6☐ Unknown/Unable to

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________________

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

D. 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** ☐

E. 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** ☐

I. Follow Up Assessment

Timing of follow up assessment: specify 8 weeks

Number of participants included in the follow up assessment: specify 126

Consistency of assessment method used: specify same method used

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: schools chosen with high proportion of Low SES and those who scored below the mean raw score (<21) on number sense screener

Specify rationale for sample size: A sample size of 126 was necessary for the planned analyses. Determined by setting α = .05 for nine predictors (age, gender, status as an English learner, nonverbal reasoning, spatial ability, attentive behavior, reading achievement, pretest score, intervention condition), with power equal to .85 and given a medium effect size of $f2$ = .15.

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 ☐

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: ______________________

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: 8 week post test follow up ____________

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: Efftects on English language learners

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 8 weeks
N

B2.2 months _____
N

B2.3 years _____
N

B2.4 other _____
N

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

C1. ☐ Unknown/insufficient information provided

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

C2.1 length of intervention session 30 minutes
N

C2.2 frequency of intervention session 3 times a week
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

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 ______

Average length of training ______

Who conducted training (select all that apply)

I2.1 ☐ Project Director
I2.2 ☐ Graduate/project assistants
I2.3 ☐ Other (please specify):
I2.3 ☐ Unknown

I3. ☐ Ongoing technical support
I4. ☐ Program materials obtained
I5. ☐ Special Facilities
I6. ☐ Other (specify):

J. Feasibility

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

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

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

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

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

Summary of Evidence for Group-Based Design Studies

Indicator	Overall Evidence Rating NNR = No numerical rating or	Description of Evidence Strong Promising Weak No/limited
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
Implementation Fidelity	2	promising
Replication	2	promising
Site of Implementation	3	strong
Follow Up Assessment Conducted	3	strong

Appendix E- Weighting of studies

The APA Task Force Coding Protocol by Kratochwill (2003) coding protocol was used to code each of the studies to generate Weight of Evidence A (WoE A). The table depicts the amendments made to the protocol, with rationale for the adaptations made.

Items excluded	Rationale
Sections 1.B.7-B.8	The studies included for review did not use qualitative research methods.
Sections II.C	The protocol was used to rate methodological quality of included studies. Outcomes are examined separately within the review.
Section II.D	All studies examined early interventions for those at risk thus were prevention programmes. Therefore this section was not relevant to the studies included in the review.
Section II.E	The intervention components are not separated in the studies reviewed.
Section II.H	The inclusion criteria required that the study be undertaken in a school therefore studies were not rated on this section.
Section III.A2	Participant characteristics for the experimental and control group are examined elsewhere in the review and it was not deemed necessary to rate methodological quality.
Section III.A3	Demographic variables and outcomes are examined elsewhere in the review.
Section III.A4	Receptivity of interventions was not deemed necessary to rate methodological quality of these studies and was not reported.

Weight of Evidence A: Methodological Quality

The score for Weight of Evidence A (WoE A) was judged based on the guidance for methodological quality provided by Kratochwill (2003) coding protocol. Studies were weighted on ‘measures’, ‘comparison group’, ‘fidelity’, ‘replication’ and ‘follow up’.

Measures

Weighting	Criteria
High – Strong evidence/support	- Studies used measures that produced reliable scores of at least 0.70 for the majority of primary outcomes. - Data collected used multiple methods and came from multiple sources. - Validity is reported.
Medium – promising evidence/support	- Studies used measures that produced reliable scores of at least 0.70. - Data collected used multiple methods and/or came from multiple sources. - A case for validity was not necessary.
Low- weak/no evidence	- Studies used measures that produced somewhat reliable scores (reliability coefficient of at least 0.50). - Data was collected using either multiple methods and/or came from multiple sources although this was not required. - A case for validity was not necessary.

Comparison group

Weighting	Criteria
High- strong evidence/support	- Studies used at least one type of active comparison group. - Initial group equivalency was established, preferably through random assignment. - Change agents were counterbalanced. - Equivalent mortality and low attrition at post and if applicable, follow up.
Medium- promising evidence/support	- At least a ‘no intervention group’ type of comparison was used. - There was at least two of the following used: counterbalancing of change agents, group equivalence or equivalent mortality with low attrition. - If equivalent mortality was not demonstrated an intent to intervene analysis was conducted.
Low – weak/no evidence	- A comparison group was used. - At least one of the following was used: counterbalancing of change agents, group equivalence was established or equivalent mortality with low attrition. - When equivalent mortality was not demonstrated an intent to intervene analysis was conducted.

Implementation fidelity

Weighting	Criteria

High- strong evidence/support	- Studies demonstrated strong evidence of acceptable adherence. - Evidence was measured through at least two of the following: ongoing supervision/consultation, coding sessions, or audio/video tapes and use of a manual. - A 'manual' was considered information provided by the implementers using either, written materials involving a detailed account of the exact procedures and sequence in which they are being used or a formal training session that included exact procedures and sequences.
Medium- promising evidence/support	- Studies demonstrated evidence of acceptable adherence. - Evidence was measured through at least one of the following: ongoing supervision/consultation, coding sessions, or audio/video tapes and use of a manual. - A 'manual' was considered information provided to the implementers using either: written materials involving an overview of broad principles and a description of the intervention phases or a formal/informal training session involving an overview of broad principles and a description of the intervention phases.
Low – weak/no evidence	- The study demonstrated evidence of acceptable adherence measured through at least one of the following: ongoing supervision/consultation, coding sessions, or audio/video tapes or use of a manual.

Replication

Weighting	Criteria
High- strong evidence/support	- The study was a replication of the same intervention and target problem - Was implemented by an independent evaluator. - Demonstrated similar or better outcomes.
Medium- promising evidence/support	- The study contained two of the coding criteria: same intervention, same target population, independent evaluation. - Demonstrated similar or better outcomes.
Low – weak/no evidence	- The study contained at least one of the following: same intervention, same target population, independent evaluation. - Demonstrated similar or better outcomes.

Follow up Assessment

Weighting	Criteria
High- strong evidence/support	- The study conducted follow up assessments over multiple intervals with all participants included in the original sample. - Similar measures were used to analyse data.
Medium- promising evidence/support	- The study must have conducted follow up assessments at least once with the majority of participants that were included in the original sample. - Similar measures were used to analyse data.
Low – weak/no evidence	- Follow up occurred at least once with some of the participants from the original sample. - No evidence indicated no follow up occurred.

Overall methodological quality

Each section was given a numerical rating from 0-3 with a score of ‘0’ indicating ‘no evidence’, ‘1’ indicating ‘low or weak evidence’, ‘2’ indicating ‘medium methodological quality with promising evidence’ and ‘3’ indicating ‘high methodological quality with strong supporting evidence’. The weightings from each section were then averaged to give an overall measure of methodological quality for the studies reviewed.

The table depicts the score equivalences which apply to all Weight of Evidence criteria.

Evidence	Score equivalences	Average Scores
Strong	High	>2.5
Promising	Medium	1.5-2.4
Weak	Low	<1.4

The table below indicates the overall weight of evidence for methodological quality (WoE A) for all the studies included in the review.

Study	Measures	Comparison group	Fidelity	Replication	Follow Up	Overall Quality of methodology
Dyson et al., (2015)	2	3	2	2	3	2.4
Gersten et al., (2015)	2	3	3	3	0	2.2
Toll and Van Luit (2012)	2	3	2	0	0	1.4
Jordan et al., (2012)	2	3	3	3	3	2.8
Bryant et al., (2011)	2	3	3	0	0	1.6
Dyson et al., (2011)	2	3	2	2	2	2.2
Aunio et al., (2005)	2	3	0	0	3	1.6

Weight of Evidence B: Methodological relevance

Weight of Evidence B (WoE B) examines the methodological relevance of the studies included in the review. WoE B considers whether the methodological design used in the studies was appropriate for evaluating the effectiveness of early intervention maths programmes for children at risk of mathematical difficulties. To decide the rating given for WoE B, an evidence hierarchy was employed (Evans, 2003) to influence descriptors. Evidence hierarchies typically place studies with high threats to validity at the bottom (for example those with no control group) and those less prone to validity at the top (for example randomised control trials). To receive a ‘high’ weighting an ‘active’ comparison group needs to be employed as this allows the effectiveness of the intervention to be examined to decide if it is better than usual or an alternative intervention. No comparison just suggests that the intervention is better than not intervening at all. Pre, post and follow up measures are necessary to receive a ‘high’ weighting as it not only provides a baseline of participants and the impact of the intervention; however it also provides information as to whether effects are maintained overtime once the intervention is removed.

Weighting	Descriptors
High	- An ‘active’ comparison group is employed. - Participants are randomly assigned to conditions. - Pre, post and follow up measures are taken for all conditions.
Medium	- A comparison group is employed e.g. wait list control/no intervention. - Participants are randomly assigned to conditions. - Pre and post measures are taken for conditions.
Low	- Pre and post outcome data is collected.

Weight of Evidence C: Topic Relevance

Weight of Evidence C (WoE C) is a review specific judgement about the relevance of the focus of the evidence for the review question.

Weighting	Descriptors
High	- The early intervention maths programme is delivered by trained professionals. - Includes children who have been identified as ‘at risk’ by some maths attainment measure. - Follow up measures are taken to see if the intervention has long term effectiveness.
Medium	- The early intervention maths programme is delivered by trained professionals. - Clear rationale is provided as to why children have been identified as ‘at risk’ for underachieving in maths and then they undertake maths attainment measures. - Pre and post measures are taken to examine the effectiveness of the intervention.
Low	- The early intervention maths programme is delivered by trained professionals. - Clear rationale is provided as to why children have been identified as ‘at risk’ for underachieving and all receive intervention. - Pre and post measures are taken to examine the effectiveness of the intervention.

Research has shown that high risk children benefit most from intensive (at least 30 minutes) specialised instruction in small groups of three to six (Gersten et al., 2007).

Therefore the delivery the maths intervention programme employs is important and should be undertaken by a trained professional to ensure the maximum effects are found. The ‘high’ weighting also considers that ‘at risk’ children must be those who are already identified as underachieving in mathematics. Whilst a clear rationale provided such as those from low SES backgrounds being more likely to be ‘at risk’ for facing maths difficulties, there may be some children who do well and achieve highly in maths and their SES does not influence their performance or achievement. Finally, if the measures are taken at pre, post and follow up these studies will receive a ‘high’ weighting as the intervention is aimed at eliminating or reducing maths difficulties in ‘at risk’ populations. Thus, long term follow up needs to be considered to ensure that intervention effects are maintained once the intervention is removed.

Weight of Evidence D: Overall Weight of Evidence

Weight of Evidence D (WoE D) is an overall assessment of the combined weightings to provide an overall appraisal of the study’s findings in relation to the review question. Using the criteria explained above, each study was given a weighting of 3 (high), 2 (medium) or 1 (low) for WoE A, B and C. These scores were then averaged to correspond to an overall weight (WoE D) for each study.

Studies	WoE A: Quality of methodology	WoE B: Relevance of methodology	WoE C: Relevance of evidence	WoE D: Overall Weight of Evidence
Dyson et al., (2015)	Medium 2.4	High 3	Medium 2	High 2.5
Gersten et al., (2015)	Medium 2.2	Medium 2	Medium 2	Medium 2.1
Toll and Van Luit (2012)	Low 1.4	Medium 2	Medium 2	Medium 1.8
Jordan et al., (2012)	High 2.8	High 3	Medium 2	High 2.6
Bryant et al., (2011)	Medium 1.6	Medium 2	Medium 2	Medium 1.9
Dyson et al., (2011)	Medium 2.2	High 3	Low 1	Medium 2.1
Aunio et al., (2005)	Medium 1.6	Medium 2	Medium 2	Medium 1.9

Studies	WoE A: Quality of methodology	WoE B: Relevance of methodology	WoE C: Relevance of evidence	WoE D: Overall Weight of Evidence

Appendix F – In depth effect size table.

Study	Measures	Number of participants	Outcomes	Effect sizes Pre-post	Effect sizes Pre- follow up	Overall Weight of Evidence
Dyson et al., (2015)	Number sense (NSB) Number list Number fact Maths achievement (WJ-III) Number list Number fact	132	Both intervention groups performed better than controls on measures of number sense and general mathematics calculation achievement at immediate post-test. However, the number-fact practice condition gave children an additional advantage over the number-list practice condition on the outcomes at delayed post-test 8 weeks later.	*g* = .32 (small) *g* = .82 (large) *g*= .58 (medium) *g* = .60 (medium)	*g* = .26 (small) *g* = .56 (medium) *g*= .12 (small) *g* =.49 (small)	High
Jordan et al., (2012)	Number Sense (NSB) Maths achievement (WJ-III)	132	Children in the number sense intervention performed better than controls at immediate post-test, with meaningful effects on measures of number sense and general math achievement. Many of the effects held at 8 week follow up.	*d* = 1.80 (large) *d* = .91 (large)	*d* = .63 (medium) *d* = .83 (large)	High

Study	Measures	Number of participants	Outcomes	Effect sizes Pre-post	Effect sizes Pre- follow up	Overall Weight of Evidence
Dyson et al., (2011)	Number Sense (NSB) Maths achievement (WJ-III)	121	The intervention group made meaningful gains relative to the control group at immediate as well delayed post-test on a measure of number sense. Intervention children also performed better than controls on maths achievement at immediate post-test.	R^2 = .30 (large) R^2 = .16 (medium)	R^2 = .29 (large)	Medium
Toll and Van Luit (2012)	Early numeracy skills (ENT-R)	192	The intervention group obtained meaningful and statistically significantly outcome numeracy scores at the post-test stage than did the control group. This was not found among the group of children falling below the 25th percentile at the pre-test stage.	η^2 = .40 (large) 25^{th}-50^{th} percentile η^2 = .22 (large)		Medium
Aunio et al., (2005)	Early numeracy skills (ENT)	45	The intervention group showed enhanced number-sense performance immediately after the instruction ended, group difference faded after six months.	η^2 = .13 (Medium) η^2 = .22 (Large)		Medium

Study	Measures	Number of participants	Outcomes	Effect sizes Pre-post	Effect sizes Pre- follow up	Overall Weight of Evidence
Gersten et al., (2015)	Mathematics ability (TEMA-3)	994	On a broad measure of mathematics ability, intervention students showed significantly greater performance compared to those in the control.	*g* = .34 (small)		Medium
Bryant et al., (2011)	Early mathematics ability (TEMI- O)	204	Students in the intervention group outperformed students in the comparison group on a broad progress-monitoring measure of mathematics performance.	*g* = .21 (small)		Medium

Note. Cohen's *d* and Hedges's *g* effect sizes are considered small at 0.2, medium at 0.5 and large at 0.8 (Cohen, 1988). Partial eta-squared (η^2) critical values are 0.01 for a small effect, 0.06 for a medium effect and 0.14 for a large effect (Cohen, 1988). R^2 represents multiple correlation coefficients squared. Values equal to .02 are considered small, those at .13 are moderate and those at .25 are considered large effects (Cohen, 1988).