International Journal of Research & Method in Education

ISSN: 1743-727X (Print) 1743-7288 (Online) Journal homepage: www.tandfonline.com/journals/cwse20

Replication in educational interventions: developing a tool to measure and promote fidelity

Terezinha Nunes, Gabriel J. Stylianides, Rosanna Lea & Louise Matthews

To cite this article: Terezinha Nunes, Gabriel J. Stylianides, Rosanna Lea & Louise Matthews (2025) Replication in educational interventions: developing a tool to measure and promote fidelity, International Journal of Research & Method in Education, 48:4, 402-423, DOI: 10.1080/1743727X.2024.2420336

To link to this article: https://doi.org/10.1080/1743727X.2024.2420336

INTERNATIONAL JOURNAL OF RESEARCH & METHOD IN EDUCATION
2025, VOL. 48, NO. 4, 402–423
https://doi.org/10.1080/1743727X.2024.2420336

Routledge
Taylor & Francis Group

Replication in educational interventions: developing a tool to measure and promote fidelity

Terezinha Nunes, Gabriel J. Stylianides, Rosanna Lea* and Louise Matthews

Department of Education, University of Oxford, Oxford, UK

ABSTRACT

The impact of many interventions weakens during the scaling up process and low fidelity of implementation (i.e. delivering an intervention but not as it was intended) may explain why. In this paper we introduce, discuss, and exemplify the use of a framework for developing fidelity tools that aim to measure and promote fidelity of implementation. The framework includes four steps that are typical of the development of new psychological and educational measures: conception, piloting and revision, implementation, and evaluation. We exemplify the use of the framework by describing the activities we carried out and the reasons for the choices we made in each stage to develop a tool to assess and promote fidelity in the scaling up of the Improving Working Memory Plus Arithmetic (IWM+A) intervention in a randomized controlled trial, involving 201 schools. In the scaling up process we used the train-the-trainer model. Twelve teacher leaders, who were trained by the programme developers, trained teaching assistants (the end users) who then delivered the intervention to children in the intervention schools (N = 100). We describe the intervention, the training model, and the process of developing, piloting, implementing, and evaluating the tool.

ARTICLE HISTORY
Received 4 April 2023
Accepted 6 September 2024

KEYWORDS
Implementation fidelity; replication; scaling up; train-the-trainer; intervention; mathematics; working memory and arithmetic

Introduction

This paper introduces a framework that can be used to develop a bespoke tool to assess intervention fidelity of educational interventions with the aim of assessing and improving fidelity of implementation. Notwithstanding the importance of replication of educational interventions, psychological and educational research face a 'replication crisis' (Perry *et al.* 2022, Schmidt and Oh 2016, Sharpe and Poets 2020). Ellefson and Oppenheimer (2023) argued that replication is not possible without *implementation fidelity*, defined in the context of educational research as 'how well an intervention is implemented in comparison with the original program design' (O'Donnell 2008, p. 34). If an intervention is not delivered as designed, participants are likely to experience something different than what was intended, producing what is known as a Type III error, i.e. measuring the outcome of something that did not take place (An *et al.* 2020, Dobson and Cook 1980). This could result in abandoning an effective intervention in the belief that it is ineffective (Kutash *et al.* 2012) or concluding that a

CONTACT Terezinha Nunes terezinha.nunes@education.ox.ac.uk Department of Education, University of Oxford, 15 Norham Gardens, Oxford, OX2 6PY, UK
*Rosanna Lea's affiliation has changed from the University of Oxford to the Education Endowment Foundation during the peer-review process.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/1743727X.2024.2420336.

certain intervention is more effective than it actually is. We recognize that fidelity of implementation is not the only threat to validity in randomized controlled trials (RCT), because lack of power also produces a major risk of reaching unjustified conclusions; in fact, lack of power can lead to failure to identify an impact as well as a magnitude error in the estimated impact, which may be over-estimated (Sims *et al.* 2022). However, a discussion of the consequences of lack of power in RCTs is beyond the scope of this paper.

It is unrealistic to expect perfect or near-perfect implementation of educational interventions at scale; perfect replication may even be unnecessary because positive results are often obtained with levels of fidelity between 60-80% (Durlak and DuPre 2008). Measuring fidelity, however, is essential to understanding why an intervention works (or not) and how its effectiveness can be optimized through identification of its 'active ingredients' (Abry *et al.* 2014, Haynes *et al.* 2016). Information about implementation fidelity can also support drawing more accurate conclusions about the effectiveness of an intervention (internal validity) and support further scaling up and replication (external validity; Ginsburg *et al.* 2021).

In educational interventions, lowering fidelity may be: (a) intentional, in order to adapt interventions to new settings (e.g. to take account of differences in budget, resources, or organizational factors; Mowbray *et al.* 2003) and to maximize efficiency and a sense of ownership (Carroll *et al.* 2007, Cordingley and Bell 2007), or (b) unintentional, resulting from specific contextual conditions (e.g. teacher knowledge; Desimone and Hill 2017), or (c) a consequence of the multifaceted nature of the intervention (e.g. Webster-Stratton *et al*. 2011). Even if an intervention has robust theoretical and empirical foundations, it may result in smaller effect sizes due to low fidelity (Borman 2007, List *et al.* 2021). A review of 500 studies (Durlak and DuPre 2008) suggested that high levels of intervention fidelity contribute to better intervention outcomes, with effect sizes 2–3 times higher for high fidelity interventions compared to those delivered with low fidelity.

The challenges in the translation of research knowledge into practice (evidence-based practice) are such that this millennium has witnessed the emergence of *implementation science*, which provides theories, models, and frameworks for understanding fidelity (Nilsen 2020, Snyder *et al.* 2013, Westerlund *et al.* 2019). Two rather different approaches to developing measures of fidelity can be distinguished: (a) a theory driven approach (e.g. An *et al.* 2020, Yeager and Walton 2011), which starts from the identification of the theoretically essential components of an intervention, or (b) an implementation approach (Outhwaite *et al.* 2020), which focuses on the barriers to, and enablers of, successful replication. Ultimately, research on fidelity should take both into account. The theory that supports the intervention provides the basis for measuring quality of implementation, whereas analysis of the context wherein an intervention is implemented provides information about the achievement of steps identified as necessary in a Theory of Change (Nelson *et al.* 2012, Reinholz and Andrews 2020). In this paper, we draw on both approaches. The theory is important as our focus is on the quality (as opposed to the dosage) of delivery and the context shaped our decisions in both developing and applying a fidelity tool to illustrate our proposed framework.

Educational interventions differ in their aims, contents, and methods of delivery. Thus, a generic measure of fidelity for evaluating different interventions may not be useful or possible. Yet a general framework can support the creation of specific tools for promoting and measuring fidelity in the context of different interventions. Drawing on the well-established tradition of psychological and educational measurement (e.g. Anastasi 1986, Cronbach 1951, Cronbach and Meehl 1956), we propose a framework for developing fidelity tools that aim to measure and promote the fidelity of implementation of different interventions. The framework has four steps: (1) conception, (2) piloting and revision, (3) implementation, and (4) evaluation. The steps are recursive (Anastasi 1986, Sechrest 1963): when the fidelity tool is used in a new project, its evaluation and revision (if required) remain important. We developed this four-step framework in the specific context of our research on the effects of an intervention on pupils' working memory and arithmetic skills. However, because the framework draws on the tradition of educational measurement, it could be used to develop tools to assess and promote the implementation fidelity of interventions on a wide range of aspects of learning.

In the next section, we describe the study background, starting from the Improving Working Memory Plus Arithmetic (IWM + A) intervention that motivated the development of our fidelity tool, and we use the project to illustrate the implementation of our proposed framework. The intervention was evaluated in a randomized controlled trial (RCT) that included an independent evaluation team, RAND Europe, appointed by the funder, the Education Endowment Foundation.[1] In the subsequent section we describe the four-step framework, the choices required at each step and the relevant literature on which we based our choices, and the activities which translated the framework into practice in the IWM + A project. We also included analyses of the reliability of the fidelity tool. In the final section we draw lessons for future research.

The research context

The intervention: improving working memory plus arithmetic (IWM + A)

The IWM + A intervention is a 10-week programme designed for 6–8 year-olds who are under-achieving in mathematics. It combines a programme for improving working memory (IWM; Nunes *et al.* 2008) with a programme for improving children's arithmetic (+A; based on Nunes *et al.* 2007, 2009). The theoretical assumption is that children's underachievement in arithmetic may be explained by domain general skills, such as Working Memory (WM) capacity, by specific arithmetic skills, or by both (Costa *et al.* 2018, Van Herwegen *et al.* 2018). Systematic reviews (Allen *et al.* 2019, Peng *et al.* 2016) have documented a robust association between WM and arithmetic skills. It is currently recognized that many school children experience cognitive-related learning difficulties (e.g. Astle *et al.* 2022, Holmes *et al.* 2019) and that there is a need for diagnostic approaches as well as specific interventions that target domain general abilities, such as working memory, and domain specific abilities, such as arithmetic skills. For children whose arithmetic difficulties are related to lower WM capacity, interventions that combine training in WM with arithmetic are more effective than those that rely on arithmetic alone (Barahmand 2008).

The design of the intervention draws on previous research findings. Performance in WM tasks depends on selective sustained attention (Johansson *et al.* 2015, Thomson *et al.* 2020), which is the ability to process parts of the sensory input to the exclusion of others and to maintain the information active for a period of time (Fisher 2019). Because WM tasks tend to be repetitive, it can be difficult for children to engage with these tasks. However, different tasks can place comparable demands on sustained attention depending on whether they form part of a game. For example, 'Track-it' (Fisher *et al.* 2013) and 'Monster Mischief' (Godwin *et al.* 2015, 2019) are selective sustained attention tasks that require children to identify the location where an object disappeared while avoiding tracking distractors. In 'Track-it', the target object is a toy that disappears behind a shape, whereas in 'Monster Mischief' the target is a monster that is said to hide behind the shape. The children's performance in the two tasks correlated significantly but, after carrying out the two tasks, the majority of the children preferred and engaged more with the 'Monster Mischief' task (Godwin *et al.* 2015).

Practicing a skill in the context of games enhances intrinsic motivation in learning and skill development (see Malone 1981, for a classic review); thus, we used games for the children to practice the use of WM and arithmetic strategies. Adult scaffolding has been found to be an effective way to improve children's attention and executive function (Hammond *et al.* 2012); a systematic review (Skene *et al.* 2022) found that guided play has a greater positive impact than direct instruction on early mathematical skills. Thus, teachers were trained in the use of scaffolding techniques to support children's learning WM and arithmetic strategies when they played the games. Finally, children are encouraged to verbalize what helps them to remember; Kloo and Perner (2003) found that children's attribution of success to the strategies they used improved learning in executive function tasks.

There is agreement that children's WM can be improved through training (e.g. see systematic review by von Bastian *et al.* 2014), but the impact of improvements in WM on educational outcomes

has been questioned. A systematic review (Melby-Lervåg and Hulme 2013) did not find convincing evidence of the generalization of WM training to other skills, such as nonverbal and verbal ability, word decoding, and arithmetic. However, none of the interventions included in the review by Melby-Lervåg and Hulme (2013) considered the role of long-term memory. The original model of WM by Baddeley and Hitch (1974) did not take long-term memory into account, but this model was revised subsequently (Baddeley 2010) to include a connection between WM and long-term memory (referred to in the model as the episodic buffer). Other models of WM included explicit connections between WM, attention and long-term memory from the outset (Cowan 1988, Ricker *et al*. 2010). The connection between WM and long-term memory can provide an explanation for the lack of transfer of gains to arithmetic tasks when a WM intervention is not associated with an intervention about arithmetic. Numbers can be represented in long-term memory in different ways. For example, the number 5 might be represented in long-term memory in isolation from other numbers; alternatively, it might be represented in connection with other numbers, if children understand additive composition (e.g. 5 can be represented as 4 + 1 or 3 + 2) and if they understand the inverse relation between addition and subtraction (e.g. 5 can be represented as 6–1 or 7-2). The amount of information that must be kept WM to connect numerical representations through additive composition and the inverse relations between addition and subtraction is considerable; consequently, children who have limited WM capacity may be at a disadvantage in developing these representations. An increase in WM resources by itself may not change number representation in long-term memory, but an intervention that combines WM and arithmetic may lead to changes in long term memory and have a positive impact children's arithmetic learning. This rationale led us to develop the intervention WM + A, which combines WM games with arithmetic games designed to improve children's understanding of additive composition.

The WM part of the intervention had been previously found effective in improving WM and attention (Nunes *et al.* 2012, 2014). The Arithmetic (A) part of the intervention also had been found effective in improving children's mathematical achievement (Worth *et al.* 2015). In a previous RCT, in which the intervention developers themselves trained the end users of the intervention, the impact of the IWM + A, which combines the WM with the A training, was compared to the impact of an intervention that used only the WM component. The children participated in the same total number of intervention sessions, but those in the WM group without the arithmetic activities only worked with WM activities. Whereas both interventions had a positive impact on WM, only the IWM + A had also a positive impact on arithmetic skills (Wright *et al.* 2019).

In the current study, the intervention developers did not train the end-users of the intervention themselves. In order to enable rolling out of the IWM + A intervention more widely, a train-the-trainer approach was adopted: the intervention developers trained trainers, who then trained the end users. To facilitate this process, it was decided to develop a tool that would enable the assessment and support of implementation fidelity.

Both parts of the IWM + A intervention consist of activities led by a Teaching Assistant (TA), delivered using pre-prepared PowerPoint slides, and games played online. In each weekly session, the TA works on a one-to-one basis with one child while, in the same room, a second child plays dedicated computer games that enable independent practice. After 30 minutes, the two children swap activities. Children should complete five WM sessions before they start the five A sessions.

During the TA-led activities, the TA teaches the use of strategies using games presented in slides. These also include reminders for the TA to prompt the children to explain their strategies and to attribute their success to the use of these strategies. In WM sessions, a 'teacher-dragon' appears on a slide with the question 'what helps you to remember?' The TA is expected to wait for the child to answer and then to display a 'student-dragon', who attributes success to the use of the strategies, consistently with the research we presented earlier (Kloo and Perner 2003). In the A sessions, slides are included to remind the TA to ask the children to explain 'why they think so' once they have answered a question. Thus, the programme contains elements that aim to promote quality of delivery, but TAs still have an important role in implementation fidelity.

In both parts of the intervention, the TA-led and the online games are designed to be adaptive and maintain the level of challenge: when children meet a criterion in one level of the game, they are moved to the next game (by the TA in the TA-led games, and automatically in the computer games). This feature has been identified as crucial for measurable impact of WM interventions (Klingberg *et al.* 2005).

The IWM + A training uses a multi-factorial approach (von Bastian and Overauer 2014), based on features previously found to be effective. Key features are explicit teaching of verbal and visuo-spatial strategies (Turley-Ames and Whitfield 2003), activities to promote focused attention training (Cowan 1988), and activities to promote number and arithmetic concepts that can affect long term memory and thus improve resource allocation (Baddeley 2010, Bruning and Lewis-Peacock 2020). In the A programme, the focus is on two concepts shown to predict and promote the development of number sense: additive composition and the inverse relation between addition and subtraction (Baroody and Lai 2007; Ching and Nunes 2017; Nunes *et al.* 2007, 2015, Schneider and Stern 2009).

Throughout the intervention, TAs are expected to teach specific strategies, to scaffold strategy use, to remove support when the children use the strategies independently, to attribute success to the use of strategies, and to give positive feedback for use of the strategies rather than to correct answers. These features are theoretically essential elements of the programme (Mowbray *et al.* 2003, Yeager and Walton 2011) and were incorporated in the design of the computerized materials as much as possible. However, TAs still have an important role to play in delivery beyond the implementation cues in the slides. In this RCT, we designed the fidelity tool to measure and promote fidelity of the teaching during the TA-led activities.

Our focus in this paper is on how we assessed and promoted the quality of delivery, which requires consideration of all the features we discussed above. Quality of delivery is crucial as an intervention can have good dosage but be delivered poorly (Carroll *et al.* 2007). Fidelity measured as quality of delivery requires a nuanced approach to capture the extent to which delivery aligns with a pre-specified standard (e.g. An *et al.* 2020; Swain *et al.* 2013).

The scaling-up and the train-the-trainer model

This paper reports on a project in which the IWM + A intervention was scaled-up using a train-the-trainer model: the programme designers trained Teacher Leaders (TLs), who then trained TAs to deliver the intervention and Link Teachers to support the TAs. RAND Europe was appointed to analyze impact and to conduct a process evaluation. Following the tradition of Theory of Change (Nelson *et al.* 2012, Reinholz and Andrews 2020), the evaluation team consulted with the programme designers to make explicit the actions and desired outcomes for this intervention and to produce a logic model, which is presented in Figure 1 (reprinted with permission; from Brown *et al.* 2021, p. 8). As in other similar research (An *et al.* 2020, Carroll *et al.* 2007, Darrow 2013, Walton *et al.* 2020), the model focuses on dosage for analysis of process and its impact on outcomes. Some information on quality of delivery was collected by the evaluation team in a small sample of schools. The evaluation team registered the protocol (Brown *et al.* 2021) and plans to publish the impact analyses in a further report.

The inputs in the yellow boxes on the left in Figure 1 describe the training of TLs (N = 12) by the programme designers, referred to in this paper as Phase 1 of the train-the-trainer model. The inputs in the green boxes as well as the activities listed in the blue boxes, referred to as Phase 2 in this paper, describe the training of TAs and Link Teachers by the TLs (i.e. the training of end users by the trainers).

TLs, who were teachers with extensive experience in training other teachers to implement interventions, participated in three one-day training sessions: an initial face-to-face familiarization session, to enable them to recruit schools for Phase 1, and two further training sessions, one on the WM component and another on the A component of the intervention. The last two sessions provided greater depth of information about the theoretical background of the intervention and the key aspects of delivery, including the opportunity to role-play delivery of TA led activities following

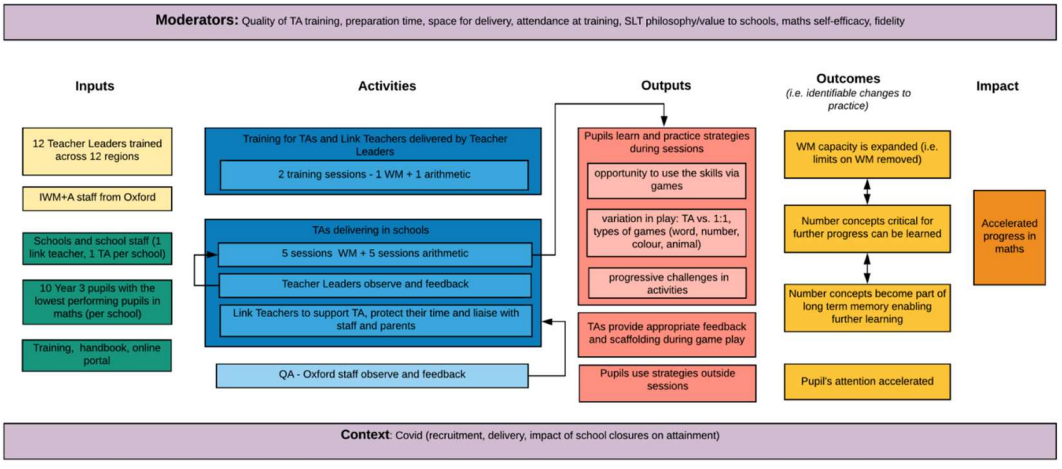

Figure 1. The Logic Model (Theory of Change) for the Improving Working Memory Plus Arithmetic intervention, developed by the independent evaluator team.

guidance from the implementation handbook as well as time to reflect and discuss what had been learned. The sessions were accompanied by a research brief that summarized key elements of the presentations.

Although our plan was to have all three training sessions with the TLs face-to-face, COVID restrictions required the latter two sessions to be virtual. Microsoft Teams was used to enable forming break-out groups for the role-play activities. After the training for delivery of each component, each TL delivered that intervention component in a school as part of their own training before going on to train TAs and Link Teachers. Each TL was observed by two researchers (the third and fourth authors) when delivering each component on one occasion. The first draft of the fidelity tool was used to collect information and to support a professional learning conversation between the researchers and the TL (see Snyder *et al.* 2013, for a similar approach).

A professional learning conversation is a discussion between professionals rather than an evaluation of teaching; for brevity, we refer to this discussion as feedback. Its purpose is to help each participant, both the observer and the end user of the intervention, to develop their ideas about teaching and learning in the specific context of the intervention. This is because a focus on outcomes (i.e. judgement about how the session went) is less effective than a focus on processes (i.e. how teaching and learning happened or can happen in this context) and can have a negative impact on learners of all ages (Dweck 2000). Thus, the aim of this feedback is to move away from observation as a judgement. To achieve this, questions and prompts are open and tentative, to enable the observer and the teacher to construct together an understanding of the observed session. This is also an opportunity for the teacher to ask questions about the intervention and to clarify any points on delivery. This approach helps teachers to: (a) build on their existing knowledge, skills and understanding; (b) articulate their thinking; (c) develop ownership of their learning; and (d) focus on the processes of teaching and learning. To support this, prompts were provided to the observer, which are presented in Figure 2. The observer's entries in the observation instrument were not shown to the TLs; rather, they were used in a careful and considerate way to guide the conversation and steer it to areas that were particularly strong or to areas that could benefit from further attention. Summaries of the outcomes of the observations were used to inform the reflection sessions on the Professional Development days.

In Phase 2, TLs recruited schools, which were randomized by the independent evaluator either to the treatment or to the control condition. In this phase, TLs used the PowerPoint slides and activities from their own training as well as videos to model delivery of the TA led activities to train TAs and Link Teachers. Training was face-to-face, whenever possible, or remotely if required by COVID-19

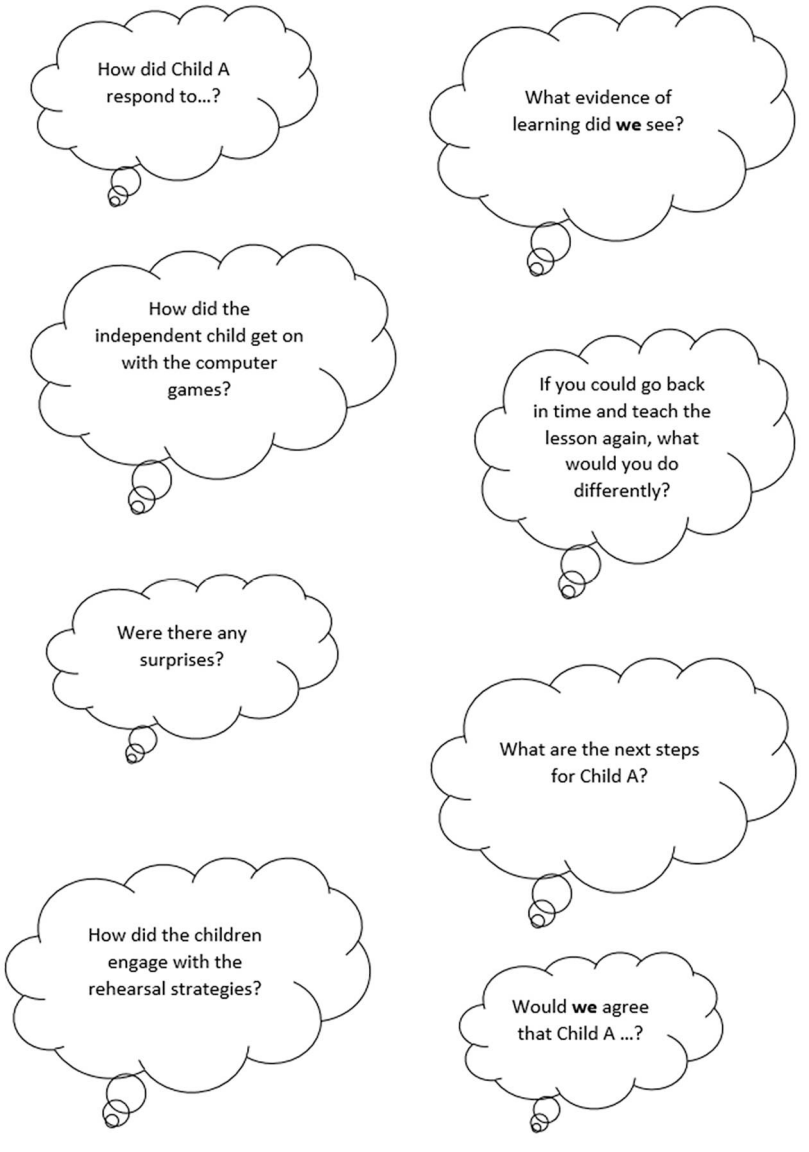

Figure 2. Prompts provided to observers for the professional learning (feedback) conversations.

restrictions. Each TL visited twice each of the TAs (with only a few exceptions as we explain in the next section) they had trained and used the revised fidelity tool to support their observations. The visit ended with a feedback session, similar to the feedback session that took place in the TL's own training. One researcher (the third or fourth authors) observed together with each TL at least once, during which the TL and the researcher used the same fidelity tool independently, in order to obtain interrater reliability data.

The framework for developing the fidelity tool and its translation into practice

In this section we will describe the four-step framework and the activities that we carried out in our project to translate the framework into practice. To develop the framework, we drew on literature on the development of psychological and educational measures (e.g. Anastasi 1986, Cronbach 1951,

Table 1. Framework for developing tools for assessing and promoting quality of delivery in (educational) interventions.

Key features		Activities
Step 1: Conception		
Choice of items	Face/content validity	Analyze documentation and consult with programme developers to list the intervention's essential elements of fidelity the tool will capture
	Evaluate consistency of list of essential elements	
	Decide on items to be used	Compare documentation analysis with a list of essential elements produced by a trained professional
		Create a list of items that correspond to the essential elements to be used in the fidelity tool, and develop clear and concise descriptors for each item, using examples where appropriate
Choice of measurement scale	Define the measurement system	Decide the format of the measurement scale, create descriptors, and explore its sensitivity to variations in delivery
	Plan how the tool will be used for promoting fidelity	Decide how the tool will be used to support a professional development conversation aimed at promoting implementation quality and fidelity
Choice of raters and reliability check	Decide who will collect data	Choose method of data collection (self-report or observation) and decide how raters will be trained
	Plan reliability check	Plan for some form of reliability information to be collected
Step 2: Piloting and revision		
Revise documentation		Revise implementation handbook to make essential elements more salient
Revise measurement		Collect a sample of observations
		Analyze scales' distributions and reliability
		Revise instructions for using the tool for scales with low reliability
Step 3: Implementation		
Consider issues about resources and guidance		Consider number of joint observations
		Plan how the tool will be used to discuss key components and fidelity
Consider practical matters		Decide when and how often fidelity will be measured
		Consider cost-effectiveness, quality assurance, and buy-in by staff delivering the intervention
Step 4: Evaluation		
Explore interrater reliability and reflect on the tool		Analyze interrater reliability and revise the tool, if required
		Reflect on what was learned by using the tool and how it can inform subsequent steps in further scaling up

Cronbach and Meehl 1956), but the step from theoretical concepts to their use in practice requires going beyond the research literature. In this paper, our description of the synergy between theoretical concepts and their use in practice provides a framework for other researchers to develop bespoke fidelity tools in the future.

In Table 1 we present an overview of the framework, which we used to develop the fidelity tool. The discussion connects the framework to relevant literature and explains how we operationalized it in the context of our study. We also present decisions required at each step as well as associated features and activities.

Step 1: Conception

Choice of items

The conception stage refers to the initial design of the fidelity tool. In order to attain face and content validity (Brandt *et al.* 2020, Yamada *et al.* 2008), consultation with the intervention designers and analysis of documentation (e.g. implementation handbook) can be used to provide an initial set of items for the tool.

In our study, a list of essential features was obtained from the trial coordinator, who was trained as a TL, in order to check the consistency of training against the implementation handbook. The revision of the documentation prompted by this analysis led to describing the theoretically essential components of the intervention as 'Keys to Success', an expression that was suggested by the TLs during this revision

process and was considered less jargony. The Keys to Success were highlighted in the implementation handbook that was later used with the TAs and in the training materials in eight bullet points; examples of the Keys to Success are: provide clear and accurate game instructions; scaffold learning; teach and remind the child of appropriate working memory strategies; play at least two games in each session; keep the level of challenge. Each 'Key to Success' heading was followed by a brief description, that was then used to create the list of items for the fidelity tool and the description of the observable elements during implementation. This led to uniformity between the description of the essential elements, the implementation handbook, and the items of the fidelity tool.

Once the fidelity tool items have been identified, attention must be paid to their wording. Clark and Watson (2019) suggested that the phrasing of an instrument influences what is ultimately measured and recommended writing principles, such as avoiding: (a) terms that may become dated quickly; (b) jargon and colloquialisms; (c) items that would apply to virtually all instances or no instances; and (d) complex or ambiguous descriptions that assess multiple characteristics. Careful consideration of phrasing may prevent later problems. For example, one of the keys to success was 'scaffold learning'. Although the word 'scaffold' is used in everyday language, its use in teaching can be considered as jargon. Because of its importance in the context of the particular intervention, rather than removing the expression, it was decided to explain it in simple terms, both in the Keys to Success and in the support for use of the fidelity scale. This led to the following description from the Keys to Success:

> You should help the child to grow their confidence in using the strategies (e.g. by rehearsing with the child initially) so that they become more independent. However, the amount of guidance and prompting required will vary between children and even between games and will generally need to be faded over time.

Choice of measurement scale

After the items have been chosen, they must be operationalized as measurement scales. For some items it might suffice to indicate whether the component is present or absent in delivery (Fogarty *et al.* 2014) and a checklist can be used. For other items, Likert scales where raters select the extent to which delivery aligns to a desired standard (e.g. 1 = not at all; 2 = somewhat; 3 = entirely) might work better. Constructing such scales requires consideration of the number of scale points and how scale options should be worded. There is no ideal number of data points for Likert scales; one often needs to make an educated guess, pilot the measure, and then add/remove options as necessary (Simms *et al.* 2019, Wakita *et al.* 2012). Scales with an odd number of points are typically used to balance the number of positive and negative points in the scale (Willits *et al.* 2016). Once the number of points in a scale is decided, labels are created to enable raters to indicate their level of agreement (e.g. from strongly disagree to strongly agree; Pence *et al.* 2008), or to assign a rating (e.g. from very poor to excellent; Lee *et al.* 2008), or to estimate how often an activity occurs (e.g. from never to very often; Cromley *et al.* 2013).

The labels should provide the opportunity for capturing what the scale is intended to measure. For example, if an essential component of the intervention is teacher provision of effective feedback, descriptors should not be related to the *quantity* of feedback (e.g. from 1 – 'Does not provide any feedback' to 5 – 'Provides lots of feedback') but to the *quality* of feedback (e.g. from 1 – 'Feedback provided is not relevant' to 5 – 'Relevant feedback is provided'). Example scenarios and ratings can offer useful guidance for training and can be compiled in a document to accompany the fidelity tool.

In our study, we decided to score all items on a 5-point scale, labelled as: 1 (Not at all), 2 (Occasionally), 3 (Sometimes), 4 (Mostly), and 5 (Always), so that higher scores reflect higher fidelity. Table 2 presents the list of items in the final version of the fidelity tool we used in Phase 2 and the corresponding descriptions.

Choice of raters and reliability check

Once the items and the measurement scale are chosen, it is necessary to decide whether self-ratings or observations will be used and what form the evaluation of reliability will take. Collecting self-

Table 2. Theoretically essential components of the IWM + A intervention.

Fidelity component	Description
1. Clarity of instruction	The TA should provide clear and accurate instructions; they should describe how games are played (as per the Handbook), and check that the child understands the rules for succeeding.
2. Teaching and reminding	The TA should teach or remind the child of appropriate strategies as necessary, and avoid teaching strategies or resources that are not part of the programme.
3. Scaffolding	The TA should scaffold learning as necessary. The child should be given support that matches their needs, using modelling, for example. Scaffolds should be reduced as children become more competent and confident, but restored if required.
4. Asking to explain	The TA should ask the child to explain their strategies at regular and helpful intervals, both when the answer is correct and when it is not.
5. Pacing	The TA should provide the child with effective opportunities to practice, at a good pace. The child should be given enough thinking time, so as not to get distracted or despondent.
6. Feedback quality	The TA should provide clear feedback on the child's use of the target strategies. The child should be praised when they use the target strategies, even if they provide incorrect answers.
7. Attributing success	The TA should explicitly attribute game success to using the target strategies, to help the child understand that the strategies are useful and will help them to succeed.
8. Level of challenge	The TA should maintain an appropriate level of challenge by ensuring that the games are played at the correct level, in the correct order. Supplementary games and extension games should be used when necessary.

reports from end users is cost-effective, but this may be biased and could overestimate fidelity (An *et al.* 2020, Toomey *et al.* 2017). Observation by trained individuals is more resource-intensive, but it is deemed the gold standard for fidelity in replication (Resnicow *et al.* 1998). Often observations and self-report data do not correlate significantly, even when the same format is used (Gresham *et al.* 2017), which complicates the decision regarding who should collect the data. Ultimately, the choice has to consider the available resources (human, financial, etc.) and the nature of what is being measured: self-reports by end users may be preferred when the aim is to measure the perception of obstacles and enablers of implementation (which can be complemented by observations by the team of intervention designers, resources permitting) (e.g. Outhwaite *et al.* 2020), whereas the quality of implementation may be evaluated more accurately by trained observers.

In our study, we opted for observation by trained individuals for a number of reasons: to obtain more accurate implementation data, to avoid burdening the end users, and to enable the feedback conversations discussed in the previous section. After deciding who will collect the data, a plan for checking reliability can be designed. In our study, we decided to have two observers attend the same session for a sample of sessions to calculate inter-rater reliability.

Step 2: Piloting and revision

Most approaches to the development of measures are iterative, involving piloting the instrument with a subset of participants and/or obtaining feedback from experts or stakeholders (Anastasi 1986, Dykstra Steinbrenner *et al.* 2015, Walton *et al.* 2020). Snyder *et al.* (2013) piloted their fidelity tool by asking trained observers to score a sample of lessons; this led to revisions of the scales used in the instrument, which has since been used in other studies (Hemmeter *et al.* 2018). Dykstra Steinbrenner *et al.* (2015) used a wider iterative approach, which coordinated the development of the intervention documentation for training alongside observations of fidelity. For example, end users of the intervention suggested changes to the implementation manual, researchers collected observations relevant to end users' ability and willingness to use the intervention, increased support and coaching was provided to improve implementation quality, and the fidelity tool was modified to capture the teacher's efforts rather than the child's behaviour. The duration and comprehensiveness of the piloting step needs to consider the nature of the intervention, including timeline, and available resources (e.g. researcher time, budget; Ginsburg *et al.* 2021).

The approach in our study was similar to that of Dykstra Steinbrenner *et al.*'s (2015), involving revision of training documentation alongside the fidelity items. Piloting and revision took place in Phase 1, which lasted approximately ten months.

Data analysis from piloting must consider the distribution of ratings and the inter-observer reliability. The distribution of ratings impacts the quality of the tool in two ways. First, a measure that does not show discrimination provides restricted information about a phenomenon. Second, if there are observer biases, these can be identified by an analysis of the distributions; distributions that depart from a normal distribution might reveal a bias in the observer's rating. However, distributions of ratings attributed to teachers trained to implement an intervention are expected to depart from a normal distribution because the teachers should be delivering the intervention at least moderately in line with the prescribed procedure. Thus, if the distribution of ratings departs from the normal distribution, this does not indicate that the rater's judgement is biased; therefore, a different approach to identifying raters' biases is required. In practice, in situations where normal distributions are not expected, it is of greater relevance to identify which aspects of the intervention seem to have been implemented with less fidelity and if there is a systematic discrepancy in the attribution of ratings between the observers (e.g. one observer uses fewer categories or uses some categories more often than another observer). The number of joint tool completions varies from study to study (e.g. 10% of all sessions in Walton *et al.* 2020) and there is no specific guidance in the literature for the proportion of joint observations required.

In our study, during Phase 1 the same two researchers from the delivery team, identified as Rater 1 and Rater 2, independently observed and coded the same intervention sessions. This procedure is considered the 'gold standard' in fidelity research (Gage *et al.* 2020, Resnicow *et al.* 1998) that aims to produce measures of inter-observer or interrater reliability (Mowbray *et al.* 2003).

Piloting in Phase 1 in our study involved observation of 11 TLs when they practiced implementation (one TL was one of the raters from the delivery team). Observations were carried via Microsoft Teams (due to COVID-19 restrictions): 6 WM sessions and 6 A sessions delivered by TLs to two children were attended by both raters, producing 11 joint observations of WM sessions (1 child was absent) and 12 joint observations of A sessions. The raters could see and hear the session and the TLs shared their screen. Whenever possible, observations took place during the third session (i.e. the mid-point) of each part of the intervention to allow the opportunity for the TA and the children to become familiar with the intervention and build rapport.

Table 3 presents information on the distribution of ratings by the two raters in Phase 1. TLs can be described as implementing all the aspects of the intervention with mid to high fidelity.

Table 3. Distribution of ratings by rater for each item and interrater reliabilities for Phase 1.

Item	Distribution of WM ratings by item by rater (rating: frequency)		Interrater reliability (κw): WM	Distribution of A ratings by item by rater (rating: frequency)		Interrater reliability (κw): Arithmetic
Average of all items	Rater 1	Rater 2	**.70**	Rater 1	Rater 2	**.64**
Clarity of instruction	3:4	3:2	.53	4:4	4:5	.43
	4:8	4:10		5:8	5:7	
Teaching and reminding	4:6	4:4	.67	4:2	4:4	.50
	5:6	5:8		5:10	5:8	
Scaffolding	4:6	4:4	.63	4:2	4:3	.83
	5:6	5:8		5:10	5:9	
Asking to explain	3:3	3:3	1.00	4:9	4:6	.43
	4:3	4:3		5:3	5:6	
	5:6	5:6				
Pacing	3:1	3:1	.63	4:9	4:8	.80
	4:3	4:6		5:3	5:4	
	5:8	5:5				
Feedback quality	3:2	3:5	.65	3:1	3:1	.50
	4:3	4:1		4:4	4:4	
	5:7	5:6		5:7	5:7	
Attributing success	3:7	3:7	.81	3:5	3:5	1.00
	4:3	4:1		4:5	4:5	
	5:2	5:4		5:2	5:2	

Notes: A = arithmetic; Rater 1 and 2 are from the delivery team rater; WM = working memory.

When the ratings were averaged across all the items, the mean rating on the scale was 4.23 (lowest score 3.29; median 4.21) for delivery of WM sessions and 4.44 (lowest score 4; median 4.36) for delivery of A sessions. Such levels of fidelity suggest that the TLs were well-prepared for intervention delivery.

The distribution of ratings by the two raters appears similar, in so far as the same range was used by them. Even though the frequency of the ratings differs, no obvious bias can be detected in either rater's attributions. Interrater reliability for each fidelity item in Phase 1 was calculated by weighted Kappa (κw). The indices per item are presented in Table 3. Interrater reliabilities were at least 'moderate' (> .41) for all 7 items; the overall scale, calculated by averaging all the ratings (consistent with de Vries *et al.* 2008), showed 'good' (> .61; Landis & Koch 1977) interrater agreement for both WM and A observations.

To improve interrater reliability in preparation for the tool's use in Phase 2, our team held detailed discussions about the definition and operationalization of all items. Items with lower interrater reliability for either WM or A sessions (i.e. clarity of instruction, teaching and reminding, asking to explain, feedback quality) were modified by including supplementary questions that asked the observer to record quantifiable behavioural indicators using a tally system before deciding which rating to attribute (Appendix A). For example, before attributing a rating for feedback quality, the rater was required first to record instances of feedback on the use of the WM strategy itself rather than on the correctness of the answer. The distinction between these two types of feedback was easier to make on an event-by-event basis rather than by rating the use of feedback at the end of the observation in a holistic way. Thus, raters were asked to record with a tick in the corresponding cell in a table whenever the child was given feedback on the use of a strategy and the strategy was discussed with the child; this allowed the rater to consider after the session if feedback had been implemented at all and, if so, with what level of consistency. Further revisions included minor changes to the wording of the tools and guidance, the addition of one item, and the re-ordering of content.

Phase 1 ratings informed the intervention designers about the need to stress two aspects in the training for delivery of the intervention, as the corresponding items had ratings below 4. Those aspects related to: (1) how to give quality feedback; and (2) how to attribute success to the use of strategies, which enhance children's motivation to use the taught strategies and improve their performance. The feedback conversations with the TLs benefitted from the use of the rating scales by helping the observers to identify target areas for discussion. The overall results were subsequently presented in the training sessions in which the TLs were prepared for training TAs in Phase 2. Emphasis was placed on the challenge to prepare TAs to give quality feedback and to be consistent in attributing success to the use of strategies during intervention delivery.

Step 3: Implementation

Implementation is the critical test of the developed fidelity tool. When a fidelity scale is used more widely, it is critical to know if it can be used with good inter-rater reliably. The reliability of a new measure refers to the degree of reproducibility attained in its repeated use. Ideally, reproducibility is assessed by repeated evaluation of the same participants by the same two (or more) raters (Gross 1986), as was done during the piloting of our fidelity measure. However, this may not be possible: for example, when a diagnostic classification (e.g. the Diagnostic Statistical Manual-III or DSM-III classification of affective disorders) is developed, a large number of clinicians must be able to use it reliably and only a small number of joint observations is feasible. It is also not possible to have the same two (or more) raters evaluating fidelity of implementation when an intervention is rolled out across a large geographical area in a large number of schools. To address this challenge, researchers working within a train-the-trainer scale-up model can, within the same training programme, train the trainers to use a fidelity scale when they observe the end users at the same time as training them how to train end users on the intervention.

This was our approach in the implementation of our fidelity tool. In our project, 100[2] treatment schools spread across 12 geographical regions in England were randomized to the intervention group. The intervention was delivered during the same two school terms. If all TAs were to be observed delivering the intervention by the same two raters from the delivery team, this would present a substantial challenge in terms of cost and capacity. Thus, we decided to train the TLs to use the fidelity tool with a dual aim: (a) to support their own feedback conversations with TAs they had trained, and (b) to help collect fidelity data on a larger scale. A sample of the sessions was observed jointly with a delivery team rater, which allowed for a comparison between the ratings by the TLs and by the delivery team rater and, consequently, for evaluation of the fidelity tool. The analyses of data from the joint observations in Phase 2 illustrate the iterative nature of the development of fidelity tools.

In Phase 2 (i.e. the implementation phase), there were 13 observers in total (two raters from the delivery team, one of whom was also a TL, plus 11 TLs). TLs visited the TAs they had trained once during a WM session and once during an A session; their ratings were used to support the feedback conversation. Whenever possible, these observations took place around the mid-point of programme delivery. One rater from the delivery team carried out joint visits with each TL to observe one WM and one A session; these were not necessarily observations of the same TA. After the feedback conversation between the TL and the TA, the delivery team rater and the TL discussed the session and their ratings; the ratings were not changed as a result of the discussion.

Rater 1 carried out joint observations with eight TLs during intervention sessions with two children, resulting in 15 joint observations (one child was absent on the day); Rater 2 carried out joint observations with four TLs, resulting in eight joint observations. Table 4 displays the descriptive statistics for each item when TAs were implementing the WM sessions. As the distributions are not normal, the median, mode, lowest (min) and highest (max) ratings are presented. The cells in sections A and B of the table are based on joint observations by TLs and raters from the delivery team.

The values in Table 4 suggest that a similar distribution of ratings was assigned by TLs and delivery team raters for the different items. The mode was lower for the item 'Attribute success to strategies' when raters were TLs, and lower for 'Scaffold' and 'Ask to explain' when the raters were from the delivery team. Across both groups of raters, the items that showed lower fidelity ratings were 'Quality of feedback' and 'Attribute success to use of strategies'.

The distribution of ratings assigned by delivery team raters across the essential features indicate that most end users delivered the intervention with good fidelity 'most of the time' (rating 4) or 'always' (rating 5); 'sometimes' (rating 3) appeared sporadically; 'occasionally' (rating 2) was assigned in 12% of the 184 ratings (23 observations x 8 scales) and 'not at all' (rating 1) was assigned in 4% of the ratings. Phase 2 data for WM sessions are in line with data from Phase 1 and suggest that end users of the intervention found it more difficult to implement feedback of the expected level of quality and to attribute the child's success to the use of the taught strategies.

Table 4. Descriptive statistics for the distributions of ratings assigned by Teacher Leaders (TLs) and by Delivery Team Raters (Rater 1 and Rater 2) for the implementation of Working Memory sessions.

	A. Descriptive statistics for ratings by TLs across Teaching Assistants (N = 23)				B. Descriptive statistics for ratings by raters from the delivery team across Teaching Assistants (N = 23)			
Item	Median	Mode	Min	Max	Median	Mode	Min	Max
Clarity	4	4	2	5	4	4	3	5
Teach	3	3	2	5	4	3	2	5
Scaffold	4	4	2	5	4	3	2	5
Ask to explain	4	4	2	5	4	3	2	5
Pacing	4	4	2	5	4	4	2	5
Quality of feedback	3	3	1	5	3	3	2	5
Attribute success to strategies	2	1	1	5	2	3	1	5
Maintain level of challenge	4	3	2	5	4	3	2	5

Table 5. Descriptive statistics for the distributions of ratings assigned by Teacher Leaders (TLs) and by Delivery Team Raters for the implementation of Arithmetic sessions.

	A. Descriptive statistics for ratings by TLs across Teaching Assistants (N = 24)				B. Descriptive statistics for ratings by Delivery Team Raters across Teaching Assistants (N = 24)			
Item	Median	Mode	Min	Max	Median	Mode	Min	Max
Clarity	4	4	2	5	4	4	2	5
Teach	4	5	2	5	4	4	3	5
Scaffold	4	4	2	5	4	4	2	5
Ask to explain	4	4	2	5	4	4	3	5
Pacing	4	4	2	5	4	4	2	5
Quality of feedback	4	4	3	5	4	4	2	5
Attribute success to strategies	3.5	4	2	5	3	3	2	5
Maintain level of challenge	4	4	2	5	4	4	2	5

A parallel analysis was carried out for the distribution of ratings by TLs and delivery team raters for the implementation of the A sessions. The descriptive statistics are presented in Table 5. The distributions of ratings attributed by both groups of raters to the items of the fidelity scale for the delivery of the A sessions were similar. The overall distribution of ratings suggests that delivery of the A sessions was very much in line with the delivery of the WM sessions. Most of the items indicate delivery with good fidelity: out of 192 ratings (24 observations x 8 ratings) by the DTRs, only 11% corresponded to 'occasionally' (rating 2) and 2% corresponded to 'not at all' (rating 1). We note that 'maintaining the level of challenge' was only included in Phase 2, as an improvement of the tool from Phase 1, and so it was not possible to compare this item with the ratings assigned by the DTRs to the sessions delivered by TLs.

Average ratings across items were calculated for each of the sessions delivered by TAs according to ratings by the delivery team raters; these are presented in Table 6. Figure 3 displays the frequency distribution for the average ratings in the fidelity scale for the WM and the A sessions. Even though measures of skewness and kurtosis do not indicate a significant departure from a normal

Table 6. Descriptive statistics for the overall ratings assigned in the fidelity scale to Teaching Assistants (TAs) by the Delivery Team Raters by type of session.

	Working Memory sessions (N = 23)	Arithmetic session (N = 24)
Mean	3.56	3.86
Median	3.50	3.94
Standard deviation	0.71	0.58
Minimum	2.50	2.75
Maximum	5.00	4.63

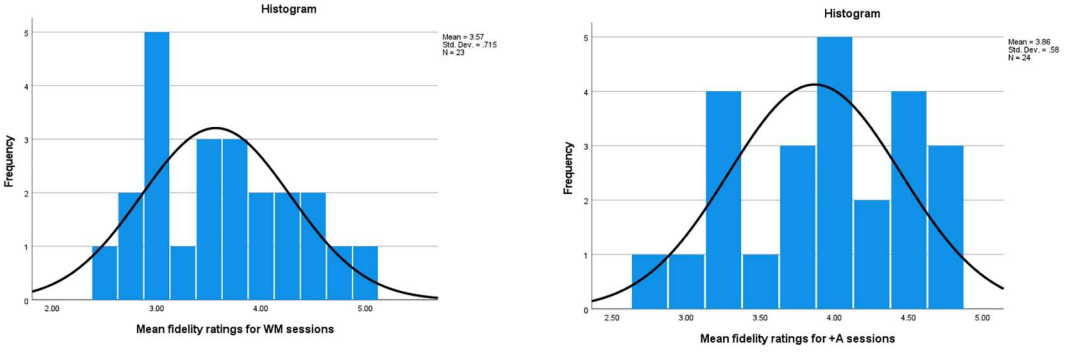

Figure 3. Mean ratings attributed to the WM and the A sessions by the Delivery Team Raters.

distribution, it can be seen in Figure 3 that both distributions are bimodal, which can be interpreted as defining two groups of sessions delivered with different levels of fidelity. We note that the TAs observed for the WM and the A sessions attended by a delivery team rater were not always the same and so we cannot tell if fidelity of delivery is correlated across type of intervention session.

The mean fidelity ratings characterizing the sessions delivered by TAs are lower than those delivered by TLs. A decrease in fidelity from sessions implemented by TLs to sessions implemented by TAs, who were the end users of the intervention, was to be expected. However, after these ratings had been attributed, there was a feedback conversation between the TL and the TA that aimed to support the TAs in improving their delivery of the intervention; it is possible that the actual fidelity of the intervention by the TAs improved subsequently.

Step 4: Evaluation

The degree of agreement between raters during Phase 1 was evaluated by the weighted Kappa, which is used when a large number of observations is collected jointly by the same two (or more) raters. However, as mentioned in the previous section, when an intervention is rolled out widely, the alternative is to use a larger number of raters and few joint observations, which means that Kappa cannot be calculated. To our knowledge, the literature does not report studies in which the inter-rater reliability of a scale was evaluated when the scale was used by a large number of raters. Thus, we searched for a novel use of statistical tools to address this problem.

According to Gross (1986), the degree of agreement in the use of a measure between the two highly trained raters (as in our Phase 1) is irrelevant to describe the level of agreement when the scale is used by a larger number of less trained raters (as in our Phase 2). Different statistical models have been proposed for describing the calibration of multiple raters (see Tong *et al.* 2020, for a review). These models are based on the assumption that the data are categorical and that there are joint observations by all the raters, even if in small numbers. However, these models are not relevant for the wide roll out of an intervention as in Phase 2 in the current study, because the measure is ordinal and joint observations by all the raters was not feasible. Therefore, a new approach to the analysis of interrater agreement was required. In these cases, models of marginal association can be used to estimate interrater agreement (Perkins and Becker 2002).

In our study, we considered a case (ordinarily a participant) to be a rating on each item for each implementation session; ratings were attributed on an ordinal scale, and each rater was defined as a variable. The two raters from the delivery team who used the scale in Phase 1 were treated as the standard against which the TLs were calibrated. A cross-tabulation for each pair of raters who observed the same sessions was run and a Kendall's tau correlation (for ordinal-by-ordinal scales) was run for the WM and the A sessions separately. Table 7 presents the observed associations between ratings by TLs and by delivery team raters for the jointly observed sessions.

It is noteworthy that, out of 24 correlations, only four were not statistically significant at $p < .01$ level; of these, two were measures of association between the ratings of the same TL and one of the delivery team raters. One can interpret these associations to suggest that there was consistency between the ratings across raters, with the exception of one rater (a TL, who might have required further training).

These results suggest that the fidelity scale developed for this project can be used with a good level of interrater agreement. It can also be used to identify raters who might require more training. If the same TLs were to be trainers in a new trial, the information from the analysis in Table 7 could be used to support a feedback conversation in preparing the TLs to train other TAs in a further roll-out of the intervention.

Reflecting on the framework, we consider its systematic use to have been of value in our study. During the conception phase, the choice of items was based on the activities suggested within the framework. As the scales were piloted, the items were revised and the need for an additional item

Table 7. Kendall's tau for ordinal-by-ordinal associations between ratings attributed by each Teacher Leaders (TL) and the Delivery Team Rater (DTR) who observed the same sessions.

TL	Associations between DTRs' and TLs' ratings for WM sessions by TL	Associations between DTRs' and TLs' ratings for A sessions by TL
1	x[a]	.62 **
2	.83**	.64**
3	.38[b]	.15[b]
4	.70**	.65**
5	.94**	.50**
6	.63**	.61**
7	.47**	.53[b]
8	.95**	.95**
9	.66**	.91**
10	.42*	.77**
11	.84**	1.0**
12	.77**	.88**

*$p < .05$; **$p < .01$.
[a]Not calculated (TL attributed 5 to all ratings in one session).
[b]ns.

was identified. Discussion of the behavioural indicators related to each item in the fidelity tool contributed to the delivery team's enhanced awareness of mediators of successful implementation. Analysis of the reliability indices during piloting led to the identification of which features seemed more difficult to implement with fidelity than others, and consequently greater emphasis on more difficult features was placed in the training of the TLs who would then train the TAs as the end users. The bimodal distribution of average fidelity ratings supported the identification of further training needs amongst the end users of the intervention. Finally, the analysis of associations between ratings assigned by the delivery team and those assigned by the TLs suggests that some TLs were more aware of the essential features than others, which identifies further training needs should the same TLs were to train other TAs in the future.

Discussion and conclusions

Although the impact of many interventions may weaken when the interventions are scaled up because of low implementation fidelity (Borman 2007, List *et al.* 2021), measures of implementation fidelity are seldom reported in educational research (Ellefson and Oppenheimer 2023). By measuring fidelity, we can understand better how and why an intervention works, support replication and scale-up efforts, and reduce the risk of Type III errors (Abry *et al.* 2014, An *et al.* 2020, Ginsburg *et al.* 2021). However, developing fidelity measures in education can be daunting, especially without a framework to guide this development process.

In this paper we introduced, discussed, and exemplified a framework that can support other researchers in the development of fidelity tools both to measure and to promote implementation fidelity. The framework, which we argue can be used in intervention research in any area of education, is based on four steps that are typical of the development of psychological and educational measures: conception, piloting and revision, implementation, and evaluation. We focused our discussion particularly on the quality of delivery (Dane and Schneider 1998), an aspect of fidelity that is more difficult to measure than other aspects (e.g. dosage) and, consequently, is often overlooked (An *et al.* 2020, Carroll *et al.* 2007, Darrow 2013, Walton *et al.* 2020).

To illustrate the utility of the framework, we reported a study that used the framework to develop a fidelity tool for the IWM + A intervention in a large RCT. In an effort to attain content validity in the tool, we started from the essential features of the IWM + A intervention, as defined by the programme developers, and identified eight theoretically essential components, which we translated into actions or behavioural indicators that could be observed during the delivery of the intervention. The items were scored using Likert scale items to provide ordinal rankings of the degree to which the

implementation approached the standard expected by the programme developers. During an initial phase, when the TLs were trained in the context of a train-the-trainer model of scaling up the intervention, the fidelity tool was piloted and used to support a conversation with a member of the core research team during which TLs reflected on their own implementation of an intervention session. TLs were subsequently trained to use the same tool both to measure fidelity and to guide similar conversations that they held with the TAs whom they trained, thereby serving the practical aim of collection fidelity data on a large scale as well as facilitating the TAs' improvement of their intervention implementation in subsequent sessions. Overall, inter-observer reliability was good when the fidelity tool was used by delivery team raters and by TLs, indicating that the tool offered a reliable means of capturing fidelity.

By detailing the process we followed to develop the fidelity tool in our study, we have illustrated how the proposed framework can be used to develop a bespoke fidelity measure which can serve as a paradigmatic case for other educational interventions and how they can use the framework to measure and promote implementation fidelity. The framework also highlights important issues and decisions that may arise throughout.

We believe that the framework may also be helpful to researchers investigating fidelity of interventions outside the field of education. We used an intervention aimed at promoting cognitive and mathematical skills to illustrate the utility of the framework, but there are no elements of the framework that are unique to this specific intervention. It is possible that different issues or decisions need to be emphasized in other contexts, but we hypothesize that the framework could be applied more widely in social science research. However, we appreciate that the utility of this framework to other projects will be dependent on a number of project-specific factors, including the resources and scope of the evaluations.

Further research might consider how to analyze the internal consistency of a fidelity scale that assesses quality of delivery. In studies that involve the measurement of an ability, such as WM or Arithmetic, the assumption is that the sources of variation in the measure include the participants' characteristics (e.g. level of skill, motivation and engagement in the task) as well as the characteristics of the items (e.g. content, form of presentation and level of difficulty). Items designed to measure fidelity of implementation can also have different sources of variation, some of which are related to the end user (the TA, in our study) of the intervention and some that are influenced by the recipient (the child). Reflecting on the types of item in our fidelity scale, we believe that items such as 'provide accurate instructions' and 'explain the strategies' may be less influenced by the child's behaviour than items which describe aspects of the implementation that are contingent on the child's behaviour, such as 'scaffold learning' or 'provide effective feedback'. In our study, we used an overall rating of fidelity of implementation because we did not have sufficient data to perform a factor analysis. However, larger studies could use factor analysis to investigate if fidelity of implementation would be best measured by using separate ratings for different dimensions.

A limitation of our study is that we were unable to analyze whether fidelity of implementation as measured by the tool is related to the impact of the IWM + A intervention. This analysis is beyond the scope of the paper and the reach of the authors, because the data to assess impact were obtained by an independent evaluation research team and not available yet to the delivery team. However, it may be possible to conduct this kind of analysis in the future, when the results become known.

To conclude, although all the steps in the framework were identified from the literature, the whole is greater than the sum of its parts. The organization of these steps into an overall framework, alongside a discussion of the issues that researchers typically face and a careful consideration of alternative choices at each step, offer a new, coherent perspective on how to approach fidelity. This overall perspective is complemented by practical descriptions of how we transformed theoretical steps into operational definitions, which can also help guide other researchers as they use the framework to develop their own bespoke tools.

Notes

1. The evaluation team was still working on its report as this paper was written.
2. Following post-randomisation and attrition of four treatment schools, the remaining 96 treatment schools were visited twice by their TL, with the exception of six schools visited only once (two due to TL sickness, and four due to TA sickness), and two were not visited at all due to long-term TA sickness.

Acknowledgements

The work reported herein received support from Education Endowment Foundation. The opinions expressed in the paper are those of the authors and do not necessarily reflect the position, policy, or endorsement of the Foundation, or of the independent evaluators, RAND Europe. The authors would like to express their gratitude to the Teacher Leaders for their commitment and enthusiasm for the project, and to staff at the schools participating in the trial for their dedication to delivering IWM + A during the COVID-19 pandemic. They also thank Peter Bryant for useful feedback on earlier versions of the paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by Education Endowment Foundation: [Grant Number].

Ethical approval

Ethical approval was granted from CUREC - Central University Research Ethics Committee in accordance with the procedures laid down by the University of Oxford for ethical approval of all research involving human participants. The approval number is ED-CIA-21-22.

ORCID

Terezinha Nunes http://orcid.org/0000-0002-1928-1805
Gabriel J. Stylianides http://orcid.org/0000-0003-1770-8753
Rosanna Lea http://orcid.org/0000-0002-2309-0948
Louise Matthews http://orcid.org/0009-0002-4465-7218

References

Abry, T., Hulleman, T., and Rimm-Kaufman, S., 2014. Using indices of fidelity to intervention core components to identify program active ingredients. *American journal of evaluation*, 36 (3), 320–338. doi:10.1177/1098214014557009.

Allen, K., Higgins, S., and Adams, J., 2019. The relationship between visuospatial working memory and mathematical performance in school-aged children: a systematic review. *Educational psychology review*, 31, 509–531. doi:10.1007/s10648-019-09470-8.

An, M., *et al.*, 2020. What really works in intervention? Using fidelity measures to support optimal outcomes. *Physical therapy*, 100 (5), 757–765. doi:10.1093/ptj/pzaa006.

Anastasi, A., 1986. Evolving concepts of test validation. *Annual review of psychology*, 37 (1), 1–16. doi:10.1146/annurev.ps.37.020186.000245.

Astle, D., *et al.*, 2022. Annual Research Review: the transdiagnostic revolution in neurodevelopmental disorders. *Journal of child psychology and psychiatry*, 63 (4), 397–417. doi:10.1111/jcpp.13481.

Baddeley, A., 2010. Working memory. *Current biology*, 20 (4), R136–R140. doi:10.1016/j.cub.2009.12.014.

Baddeley, A.D., and Hitch G.J., 1974. Working memory. In: G.H. Bower, ed. *The psychology of learning and motivation*. Academic Press, 47–89.

Barahmand, U., 2008. Arithmetic disabilities: training in attention and memory enhances arithmetic ability. *Research journal of biological sciences*, 3 (11), 1305–1312.

Baroody, A., and Lai, M., 2007. Preschoolers' understanding of the addition–subtraction inverse principle: a Taiwanese sample. *Mathematical thinking and learning*, 9 (2), 131–171. doi:10.1080/10986060709336813.

Borman, G., 2007. National efforts to bring reform to scale in high-poverty schools: outcomes and implications. In: B. Schneider, and S.K. McDonald, eds. *Scale-up in education: issues in practice*. Rowan and Littlefield Publishers, 41–67.

Brandt, N., McHale, C., and Humphris, G., 2020. Development and testing of a novel measure to assess fidelity of implementation: example of the mini-AFTERc intervention. *Frontiers in psychology*, 11, 601813.

Brown, E., *et al*., 2021. *Improving Working Memory Plus Arithmetic evaluation protocol*. Education Endowment Foundation. Available from: https://d2tic4wvo1iusb.cloudfront.net/documents/pages/projects/IWMA-protocol-final.pdf?v=1680208530.

Bruning, A., and Lewis-Peacock, J., 2020. Long-term memory guides resource allocation in working memory. *Scientific reports*, 10 (1), 1–10. doi:10.1038/s41598-020-79108-1.

Carroll, C., *et al*., 2007. A conceptual framework for implementation fidelity. *Implementation science*, 2 (40): 1–9. doi:10.1186/1748-5908-2-40.

Ching, B.H.H., and Nunes, T., 2017. The importance of additive reasoning in children's mathematical achievement: a longitudinal study. *Journal of educational psychology*, 109 (4), 477. doi:10.1037/edu0000154.

Clark, L.A., and Watson, D., 2019. Constructing validity: new developments in creating objective measuring instruments. *Psychological assessment*, 31 ((12), 1412–1427. doi:10.1037/pas0000626.

Cordingley, P., and Bell, M. 2007. *Transferring learning and taking innovation to scale*. Centre for the Use of Research and Evidence in Education. Available from: http://curee.co.uk/resources/publications/transferring-learning-and-taking-innovation-scale.

Costa, H., *et al*., 2018. Low performance on mathematical tasks in preschoolers: the importance of domain-general and domain-specific abilities. *Journal of intellectual disability research*, 62 (4), 292–302. doi:10.1111/jir.12465.

Cowan, N., 1988. Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological bulletin*, 104 (2), 163–191. doi:10.1037/0033-2909.104.2.163.

Cromley, J., *et al*., 2013. Improving students' diagram comprehension with classroom instruction. *The journal of experimental education*, 81 (4), 511–537. doi:10.1080/00220973.2012.745465.

Cronbach, L., 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.

Cronbach, L., and Meehl, P., 1956. Construct validity in psychological tests. *Minnesota psychological bulletin*, 52 (4), 281–302. doi:10.1037/h0040957.

Dane, A., and Schneider, B., 1998. Program integrity in primary and early secondary prevention: are implementation effects out of control? *Clinical psychology review*, 18 (1), 23–45. doi:10.1016/S0272-7358(97)00043-3.

Darrow, C., 2013. The effectiveness and precision of intervention fidelity measures in preschool intervention research. *Early education & development*, 24, 1137–1160. doi:10.1080/10409289.2013.765786.

Desimone, L., and Hill, K., 2017. Inside the black box: examining mediators and moderators of a middle school science intervention. *Educational evaluation and policy analysis*, 39 (3), 511–536. doi:10.3102/0162373717697842.

De Vries, H., *et al*., 2008. Using pooled kappa to summarize interrater agreement across many items. *Field methods*, 20 (3), 272–282. doi:10.1177/1525822X08317166.

Dobson, D., and Cook, T., 1980. Avoiding type III error in program evaluation: results from a field experiment. *Evaluation and program planning*, 3 (4), 269–276. doi:10.1016/0149-7189(80)90042-7.

Durlak, J., and DuPre, E., 2008. Implementation matters: a review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American journal of community psychology*, 41, 327–350. doi:10.1007/s10464-008-9165-0.

Dweck, C., 2000. *Self-theories: their role in motivation, personality and development*. Hove: Brunner/Mazel.

Dykstra Steinbrenner, J., *et al*., 2015. Developing feasible and effective school-based interventions for children with ASD: a case study of the iterative development process. *Journal of early intervention*, 37 (1): 23–43. doi:10.1177/1053815115588827.

Ellefson, M., and Oppenheimer, D., 2023. Is replication possible without fidelity? *Psychological methods* 26 (6): 1446–1452. doi:10.1037/met0000473.

Fisher, A., *et al*., 2013. Assessing selective sustained attention in 3- to 5-year-old children: evidence from a new paradigm. *Journal of experimental child psychology*, 114 (2), 275–294. doi:10.1016/j.jecp.2012.07.006.

Fisher, A., 2019. Selective sustained attention: a developmental foundation for cognition. *Current opinion in psychology*, 29, 248–253. doi:10.1016/j.copsyc.2019.06.002.

Fogarty, M., *et al*., 2014. Examining the effectiveness of a multicomponent reading comprehension intervention in middle schools: a focus on treatment fidelity. *Educational psychology review*, 26, 425–449. doi:10.1007/s10648-014-9270-6.

Gage, N., MacSuga-Gage, A., and Detrich, R., 2020. *Fidelity of implementation in Educational Research and Practice*. Oakland, CA: The Wing Institute. https://www.winginstitute.org/systems-program-fidelity.

Ginsburg, L., *et al*., 2021. Fidelity is not easy! Challenges and guidelines for assessing fidelity in complex interventions. *Trials*, 29 (22), 371. doi:10.1186/s13063-021-05322-5.

Godwin, K., *et al*., 2015. Monster Mischief: designing a video game to assess selective sustained attention. *International journal of gaming and computer-mediated simulations*, 7 (4), 18–39. doi:10.4018/IJGCMS.2015100102.

Godwin, K., *et al*., 2019. Monster Mischief: a game-based assessment of selective sustained attention in young children. In: *Exploring the cognitive, social, cultural, and psychological aspects of gaming and simulations*. IGI Global, 171–205.

Gresham, F., Dart, E., and Collins, T., 2017. Generalizability of multiple measures of treatment integrity: comparisons among direct observations, permanent products, and self-report. *School psychology review*, 46 (1), 108–121. doi:10.1080/02796015.2017.12087606.

Gross, S.T., 1986. The kappa coefficient of agreement for multiple observers when the number of subjects is small. *Biometrics*, 883–893. doi:10.2307/2530702.

Hammond, S., *et al*., 2012. The effects of parental scaffolding on preschoolers' executive function. *Developmental psychology*, 48 (1), 271–281. doi:10.1037/a0025519.

Haynes, A., *et al*., 2016. Figuring out fidelity: a worked example of the methods used to identify, critique and revise the essential elements of a contextualised intervention in health policy agencies. *Implementation science*, 11 (23): 1–18. doi:10.1186/s13012-016-0378-6.

Hemmeter, M.L., Snyder, P., and Fox, L., 2018. Using the Teaching Pyramid Observation Tool (TPOT) to support implementation of social-emotional teaching practices. *School mental health*, 10, 202–213. doi:10.1007/s12310-017-9239-y.

Holmes, J., Bryant, A., CALM team, and Gathercole, S., 2019. Protocol for a transdiagnostic study of children with problems of attention, learning and memory (CALM). *BMC pediatrics*, 19 (1), 10. doi:10.1186/s12887-018-1385-3.

Johansson, M., *et al*., 2015. Sustained attention in infancy as a longitudinal predictor of self-regulatory functions. *Infant behavior and development*, 41, 1–11.

Klingberg, T., *et al*., 2005. Computerized training of working memory in children with ADHD-a randomized, controlled trial. *Journal of the American academy of child & adolescent psychiatry*, 44, 177–186. doi:10.1097/00004583-200502000-00010.

Kloo, D., and Perner, J., 2003. Training transfer between card sorting and false belief understanding: helping children apply conflicting descriptions. *Child development*, 74 (6), 1823–1839. doi:10.1046/j.1467-8624.2003.00640.x.

Kutash, K., *et al*., 2012. Description of a fidelity implementation system: an example from a community-based children's mental health program. *Journal of child and family studies*, 21 (6), 1028–1040. doi:10.1007/s10826-012-9565-5.

Landis, J.R., and Koch, G., 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33 (1), 159–174. doi:10.2307/2529310.

Lee, C.Y., *et al*., 2008. Fidelity at a distance: assessing implementation fidelity of the Early Risers Prevention Program in a going-to-scale intervention trial. *Prevention science*, 9 (3), 215–229. doi:10.1007/s11121-008-0097-6.

List, J., Suskind, D., and Supplee, L., 2021. *The scale-up effect in early childhood and public policy: why interventions lose impact at scale and what we can do about it*. New York: Routledge.

Malone, T., 1981. Toward a theory of intrinsically motivating instruction. *Cognitive science*, 5 (4), 333–369. doi:10.1207/s15516709cog0504_2.

Melby-Lervåg, M., and Hulme, C., 2013. Is working memory training effective? A meta-analytic review. *Developmental Psychology*, 49 (2), 270–291. doi:10.1037/a0028228.

Mowbray, C., *et al*., 2003. Fidelity criteria: development, measurement and validation. *American journal of evaluation*, 24 (3), 315–340. doi:10.1177/109821400302400303.

Nelson, M., *et al*., 2012. A procedure for assessing intervention fidelity in experiments testing educational and behavioral interventions. *The journal of behavioral health services & research*, 39 (4): 374–396. doi:10.1007/s11414-012-9295-x.

Nilsen, P., 2020. Overview of theories, models and frameworks in implementation science. In: P. Nilsen, and S. Birken, eds. *Handbook on implementation science*. Cheltenham, UK: Edward Elgar Publishing, 8–31.

Nunes, T., *et al*., 2007. The contribution of logical reasoning to the learning of mathematics in primary school. *British journal of developmental psychology*, 25, 147–166. doi:10.1348/026151006X153127.

Nunes, T., *et al*., 2008. *Improving children's working memory through guided rehearsal*. New York: American Educational Research Association (AERA).

Nunes, T., *et al*., 2009. Teaching children about the inverse relation between addition and subtraction. *Mathematical thinking and learning*, 11 (1), 61–78. doi:10.1080/10986060802583980.

Nunes, T., *et al*., 2012. A game-based working memory intervention for deaf children. In: S. Wannemacker, S. Vandercruyse, and G. Clarebout, eds. *Serious games: the challenge*. Berlin: Springer-Verlage, 31–39.

Nunes, T., *et al*., 2014. Improving Deaf children's working memory through training. *International journal of speech & language pathology and audiology*, 2, 51–66. doi:10.12970/2311-1917.2014.02.02.1.

Nunes, T., *et al*., 2015. Assessing quantitative reasoning in young children. *Mathematical thinking and learning*, 17 (2-3), 178–196. doi:10.1080/10986065.2015.1016815.

O'Donnell, C., 2008. Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. *Review of educational research*, 78 (1), 33–84. doi:10.3102/0034654307313793.

Outhwaite, L., Gulliford, A., and Pitchford, N., 2020. A new methodological approach for evaluating the impact of educational intervention implementation on learning outcomes. *International journal of research & method in education*, 43 (3), 225–242. doi:10.1080/1743727X.2019.1657081.

Pence, K., Justice, L., and Wiggins, A., 2008. Preschool teachers' fidelity in implementing a comprehensive language-rich curriculum. *Language, speech, and hearing services in schools*, 39, 329–341. doi:10.1044/0161-1461(2008/031).

Peng, P., *et al.*, 2016. A meta-analysis of mathematics and working memory: moderating effects of working memory domain, type of mathematics skill, and sample characteristics. *Journal of educational psychology*, 108 (4), 455–473. doi:10.1037/edu0000079.

Perkins, S., and Becker, M., 2002. Assessing rater agreement using marginal association models. *Statistics in medicine*, 21 (12), 1743–1760. doi:10.1002/sim.1146.

Perry, T., Morris, R., and Lea, R., 2022. A decade of replication study in education? A mapping review (2011-2020). *Educational research and evaluation*, 27, 12–34. doi:10.1080/13803611.2021.2022315.

Reinholz, D., and Andrews, T., 2020. Change theory and theory of change: what's the difference anyway? *International journal of STEM education*, 7 (2): 1–12. doi:10.1186/s40594-020-0202-3.

Resnicow, K., *et al.*, 1998. How best to measure implementation of school health curricula: a comparison of three measures. *Health education research*, 13 (2), 239–250. doi:10.1093/her/13.2.239.

Ricker, T.J., Aubuchon, A.M., and Cowan, N., 2010. Working memory. *Wiley interdisciplinary reviews: Cognitive science* 1 (4): 573–585.

Schmidt, F., and Oh, I.S., 2016. The crisis of confidence in research findings in psychology: is lack of replication the real problem? Or is it something else? *Archives of scientific psychology*, 4 (1), 32–37. doi:10.1037/arc0000029.

Schneider, M., and Stern, E., 2009. The inverse relation of addition and subtraction: a knowledge integration perspective. *Mathematical thinking and learning*, 11 (1–2), 92–101. doi:10.1080/10986060802584012.

Sechrest, L., 1963. Incremental validity: a recommendation. *Educational and psychological measurement*, 23 (1), 153–158. doi:10.1177/001316446302300113.

Sharpe, D., and Poets, S., 2020. Meta-analysis as a response to the replication crisis. *Canadian psychology / Psychologie canadienne*, 61 (4), 377–387. doi:10.1037/cap0000215.

Simms, L., *et al.*, 2019. Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological assessment*, 31 (4), 557–566. doi:10.1037/pas0000648.

Sims, S., *et al.*, 2022. Quantifying "promising trials bias" in randomized controlled trials in education. *Journal of research on educational effectiveness* 16 (4): 663–680. doi:10.1080/19345747.2022.2090470.

Skene, K., *et al.*, 2022. Can guidance during play enhance children's learning and development in educational contexts? A systematic review and meta-analysis. *Child development*, 93 (4), 1162–1180. doi:10.1111/cdev.13730.

Snyder, P., *et al.*, 2013. Developing and gathering psychometric evidence for a fidelity instrument: the Teaching Pyramid Observation Tool-Pilot Version. *Journal of early intervention*, 35 (2), 150–172. doi:10.1177/1053815113516794.

Swain, M., Finney, S., and Gerstner, J., 2013. A practical approach to assessing implementation fidelity. *Assessment update*, 25 (1): 5–7. doi:10.1002/au.251.

Thomson, P., *et al.*, 2020. Longitudinal trajectories of sustained attention development in children and adolescents with ADHD. *Journal of abnormal child psychology*, 48 (12), 1529–1542. doi:10.1007/s10802-020-00698-5.

Tong, F., *et al.*, 2020. The determination of appropriate coefficient indices for inter-rater reliability: using classroom observation instruments as fidelity measures in large-scale randomized research. *International journal of educational research*, 99, 101514.

Toomey, E., Matthews, J., and Hurley, D., 2017. Using mixed methods to assess fidelity of delivery and its influencing factors in a complex self-management intervention for people with osteoarthritis and low back pain. *BMJ open*, 77 (8): e015452. doi:10.1136/bmjopen-2016-015452.

Turley-Ames, K., and Whitfield, M., 2003. Working memory training and task performance. *Journal of memory and language*, 49, 446–468. doi:10.1016/S0749-596X(03)00095-0.

Van Herwegen, J., *et al.*, 2018. Improving number abilities in low achieving preschoolers: symbolic versus non-symbolic training programs. *Research in developmental disabilities*, 77, 1–11. doi:10.1016/j.ridd.2018.03.011.

Von Bastian, C., and Oberauer, K., 2014. Effects and mechanisms of working memory training: a review. *Psychological research*, 78 (6), 803–820. doi:10.1007/s00426-013-0524-6.

Wakita, T., Ueshima, N., and Noguchi, H., 2012. Psychological distance between categories in the Likert scale: comparing different numbers of options. *Educational and psychological measurement*, 72 (4), 533–546. doi:10.1177/0013164411431162.

Walton, H., *et al.*, 2020. Developing quality fidelity and engagement measures for complex health interventions. *British journal of health psychology*, 25 (1), 39–60. doi:10.1111/bjhp.12394.

Webster-Stratton, C., *et al.*, 2011. The incredible years teacher classroom management training: the methods and principles that support fidelity of training delivery. *School psychology review*, 40 (4), 509–529.

Westerlund, A., Sundberg, L., and Nilsen, P., 2019. Implementation of implementation science knowledge: the research-practice gap paradox. *Worldviews on evidence-based nursing*, 16 (5), 332–334. doi:10.1111/wvn.12403.

Willits, F., Theodori, G., and Luloff, A.E., 2016. Another look at Likert scales. *Journal of rural social sciences*, 31 (3), 6. https://egrove.olemiss.edu/jrss/vol31/iss3/6.

Worth, J., *et al.* 2015. *Improving numeracy and literacy: evaluation report and executive summary*. Education Endowment Foundation. Available from: https://www.nfer.ac.uk/media/1690/eeol01.pdf.

Wright, H., *et al.* 2019. *Improving numeracy and literacy: evaluation report*. Education Endowment Foundation. Available from: https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/improving-working-memory.

Yamada, J., *et al.*, 2008. Validation of a process evaluation checklist to measure intervention implementation fidelity. *Archives of disease in childhood*, 93 (Suppl 2), 289.

Yeager, D., and Walton, G., 2011. Social-psychological interventions in education: they're not magic. *Review of educational research*, 81 (2), 267–301. doi:10.3102/0034654311405999.