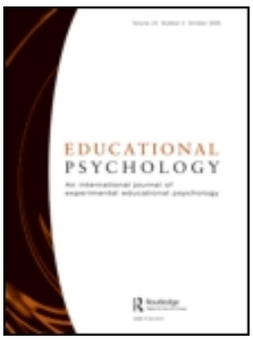

Routledge
Taylor & Francis Group

Educational Psychology

An International Journal of Experimental Educational Psychology

ISSN: 0144-3410 (Print) 1469-5820 (Online) Journal homepage: https://www.tandfonline.com/loi/cedp20

Implementation integrity in peer tutoring of mathematics

Keith Topping , David Miller , Pauline Murray & Nora Conlin

To cite this article: Keith Topping , David Miller , Pauline Murray & Nora Conlin (2011) Implementation integrity in peer tutoring of mathematics, Educational Psychology, 31:5, 575-593, DOI: 10.1080/01443410.2011.585949

To link to this article: https://doi.org/10.1080/01443410.2011.585949

Educational Psychology
Vol. 31, No. 5, August 2011, 575–593

Routledge
Taylor & Francis Group

Implementation integrity in peer tutoring of mathematics

Keith Topping[a]*, David Miller[b], Pauline Murray[c] and Nora Conlin[d]

[a]*School of Education, University of Dundee, Dundee, UK;* [b]*University of Dundee, Dundee, UK;* [c]*Angus Council, Forfar, UK;* [d]*Fife Council, Dunfermline, UK*

(*Received 28 September 2010; final version received 2 May 2011*)

A two-year randomised controlled trial of peer tutoring in mathematics using the Duolog Math technique was operated in 80 schools. The aim was to achieve adequate implementation quality with modest pre-intervention training for teachers, who received brief didactic training and no process feedback (but they were to train pupils using modelling, practice and feedback). Implementation integrity was observed in Year 1 in 29 randomly selected schools; in Year 2 in 30 randomly selected schools. In both years some observed variables of class-wide context and individual technique were high; however, some were lower. There were deficits in: teachers introducing the problem, suggesting ways to concretise the problem and holding plenary sessions. Crucially, there was very little summarising or generalising. Thus, implementation was partial in both years, but better in Year 2. The implications for future intensity of training are explored.

Keywords: mathematics; peer tutoring; implementation integrity; implementation fidelity; observation

The aim of this study was to see what quality of implementation integrity was achievable with modest pre-intervention training for teachers. The context was a two-year randomised controlled trial of peer tutoring of mathematics involving 80 schools in one school district, using the Duolog Math technique. Schools were assigned mathematics or reading + mathematics tutoring, light or intensive intervention and same-age or cross-age tutoring. Undertaking process measurement of all 80 schools would be an enormous task, so in this study, implementation quality was assessed by direct observation by researchers in 29 and 30 randomly selected schools in each year (without control group measurement). The quality of implementation integrity of math tutoring (in relation to the relatively brief and simple initial training for teachers) was likely to be crucial in the successfulness of the intervention. In this study, the implementation quality of the teacher's initial attempt to train the children was not measured. What was measured was the quality of class-wide teacher–pupil supportive interaction and the quality of tutor–tutee technique at mid-point. Research questions consequently concerned the level of implementation integrity, difference between years and whether general contextual classroom factors were associated with the quality of individual tutoring technique. This study focuses on process.

*Corresponding author. Email: k.j.topping@dundee.ac.uk

ISSN 0144-3410 print/ISSN 1469-5820 online

DOI: 10.1080/01443410.2011.585949
http://www.informaworld.com

Background

This section discusses the nature of peer tutoring. It then discusses implementation integrity and connects this with quality of pre-intervention training. Both are key to the aim of this study. It then discusses the tutoring technique used here, and research on its effectiveness, in relation to the structure of the process observations.

Peer tutoring

The current intervention was peer tutoring – characterised by specific role-taking (tutor or tutee), and usually one-on-one tuition and high focus on curriculum content. There are specific procedures for interaction, in which participants are likely to have training that is generic or specific or both. The effects of social interactivity, discourse, reflection and motivation feature in theoretical models of tutoring (e.g. Topping & Ehly, 1998). The literature is large and studies tend to show benefits both for the tutors – perhaps due to their need to reflect and communicate concepts and procedures – and for the tutees (Britz, Dixon, & McLaughlin, 1989; Cohen, Kulik, & Kulik, 1982; Robinson, Schofield, & Steers-Wentzell, 2005; Rohrbeck, Ginsburg-Block, Fantuzzo, & Miller, 2003). In a review, Ginsburg-Block, Rohrbeck, and Fantuzzo (2006) found that structured pupil roles within which pupils had a degree of personal autonomy, individualised evaluation, interdependent group rewards and same-gender grouping accounted for most of the impact. The tutoring technique used in the current study featured highly structured pupil roles and interaction and a high degree of pupil autonomy and individual evaluation, but did not feature interdependent group rewards or same-gender matching (the first was hypothesised as a complication of uncertain worth in the UK context, and the second potentially restrictive in classes with different numbers of males and females).

Tutoring procedures are most effective when thoroughly scaffolded (Cohen et al., 1982; Sharpley & Sharpley, 1981; Topping & Ehly, 1998), since untrained tutoring behaviours tend to be primitive (e.g. Person & Graesser, 1999), characterised by infrequent correction of errors and inappropriate giving of positive feedback. Roscoe and Chi (2007) suggested that analysis of actual behaviours was important in enhancing the effectiveness of tutoring. Explaining and questioning were important in reflective knowledge-building, but tutors tended to persist with simple knowledge-telling.

Implementation integrity

Implementation integrity (or implementation fidelity) is the extent to which an intervention is delivered according to the description of good practice in the technical protocol or manual for it (Arkoosh et al., 2007; Gresham, 1989). Gresham argued that implementation integrity concerned descriptive features of the programme (complexity of treatments, time required to implement, materials/resources, number of treatment agents) and technical issues (specification of treatment components and deviations from treatment protocols), and are best determined by direct observation. Arkoosh et al. (2007) expressed this in more relative terms – what are the parameters required to achieve *desired* treatment outcomes – how much integrity is

enough? Thus, implementation integrity might constitute a continuum which relates to a continuum in outcomes (and indeed in implementation cost). These authors found that a high level of treatment integrity was required for treatment success. However, implementation of *all* elements of a treatment plan may not be critical to such success (Gansle & McMahon, 1997).

Generally, higher implementation integrity is associated with higher outcomes (e.g. Noell et al., 2005), but this relationship is not necessarily strong or continuous. These authors note that intervention implementation was typically initially high, but decreased over time, and in many instances the deterioration was precipitous. Clearly, a purely outcome evaluation with positive or negative results would not be able to claim that the target programme had actually been evaluated if there was no evidence of implementation integrity. This is a critical issue for most response-to-intervention models. However, there has been relatively little research on implementation integrity, particularly in large-scale field trials. In schools, implementation integrity can focus on teacher behaviour, or pupil behaviour, or both (the latter approach is used here). Of course, the more complex the procedure, the more demanding is any assessment of implementation integrity.

Nature of pre-intervention training and subsequent implementation

The extent to which the brief pre-service training without systematic feedback in this study was sufficient to change the behaviour of the teachers raises interesting questions. Duolog Math was new to the teachers, so considerable behaviour change was required. Gilbertson, Witt, Singletary, and VanDerHeyden (2007) gave verbal instruction to five teachers, after which implementation was consistently low or exhibited a downward trend. Three subsequent classroom training sessions improved implementation quality, but obviously at a cost in terms of time. Response-dependent feedback was necessary to raise implementation in all teachers to the required level. We hoped to be more effective than this in the present study.

The nature of training is important and easier with instructional standardisation, which can generate higher levels of instructional change in the longer term (Rowan & Miller, 2007). However, these authors found not all implementations were perfect. Blom-Hoffman (2007) emphasised the importance of involving many team members in implementation, each taking an element of responsibility, so that management was devolved and many players felt engaged in the process, but in a time-efficient manner. In a study of 10 elementary schools, Taylor and Teddlie (1999) found that that partial implementation was generally the rule, with more challenging innovations the least likely to be implemented. Teachers exhibit fidelity to activity more readily than to instructional processes (Pence, Justice, & Wiggins, 2008) – they implement the mechanics of what to do better than the subtleties of why these things need to be done. These authors found teacher use of targeted instructional processes relatively low even after a year of implementation. Stead, Stradling, Macneil, Mackintosh, and Minty (2007) found that intervention sessions were delivered with reasonable fidelity, but the teachers did not always understand the thinking behind particular activities, suggesting that training needed to focus not only on content and methods but also on why particular approaches were important. Under pressure of time, generic elements and processes designed to reflect on learning were sometimes sacrificed in order that core activities could be completed.

Time is an important feature in relation to measurement, since after training teachers might implement mechanics first and develop greater insight into process subtleties over time, but process observation might not occur sufficiently frequently to capture this. Some aspects of interventions may be implemented faster than others (Bradshaw, Reinke, Brown, Bevans, & Leaf, 2008). Additionally, trained schools can show higher implementation integrity, but untrained schools also show increases. Teachers will usually make some change to programmes, both in content and in method, and most of these modifications attenuate programme effects. Consequently, there should be particular emphasis on the key features of programmes, in order that the core might be better implemented (Pankratz et al., 2006).

The extent of performance feedback in pre-intervention training might be an important variable in subsequent implementation quality. Auld, Belfiore, and Scheeler (2010) used brief initial training coupled with individual feedback meetings following direct classroom observation to change the behaviour of seven teachers with respect to the use of differential reinforcement. Sterling-Turner, Watson, Wildmon, Watkins, and Little (2001) found modelling and rehearsal with feedback more effective than purely instructor-directed methods in establishing treatment integrity. Noell et al. (2005) found that performance feedback was associated with superior treatment implementation and child behavioural outcomes when compared to other conditions. Performance feedback could be used as an efficient means of improving or maintaining treatment integrity when applied within a team model (Duhon, Mesmer, Gregerson, & Witt, 2009). Burns, Peters, and Noell (2008) provided structured performance feedback in an attempt to enhance integrity, and found a short-term immediate change, but in the longer term, significant failures were still evident. Teacher goal-setting with written performance feedback about pupil performance was less effective than feedback about teacher and pupil performance with a contingency of avoiding practicing missed steps (Digennaro, Martens, & Kleinmann, 2007). However, these suggestions for extended feedback all imply extra input time.

Treatment – the tutoring technique

Duolog Math uses discussion between peers to help solve the problem (hence the name). Tutoring is always cross-ability, but peer matching aims for a similar and modest differential in ability in all pairs. Too large or small an ability differential might restrict attainment gain in both members of the pair (Topping & Ehly, 1998). To facilitate the discussion, nine interactive behaviours are proposed: Read, Listen, Check, Praise; Pause for Think-Aloud, Question, Make it Real, Summarise and Generalise. These behaviours were selected in accordance with a theoretical model (Topping & Ehly, 1998) as suitable for peer tutoring from a much longer listing of math coaching behaviours undertaken by math teachers. Normally, teachers train tutors and tutees in the relatively easy first four steps together. The second four steps are considerably harder, and teachers often introduce these one week at a time, devoting a mini-lesson to each.

A description of each behaviour is given below in the language used in the project:

(1) Read – your tutee might be having trouble reading a word problem. If so, read it for them and check their understanding.

(2) Listen – give your tutee time to struggle to explain what his/her difficulty is. Do not just jump in to fix what you assume the problem is.
(3) Check – check that your tutee eventually gets the right answer. But remember there is probably more than one 'right' way to solve the problem. If the answer is wrong read the problem over and try again. Only if all else fails show your tutee how you would do it (while you think aloud).
(4) Praise and Encourage – give your tutee praise and encouragement often, even for a very small success with a single step in solving a problem. Keep their confidence high.
(5) Pause for Think-Aloud – give your tutee some thinking time, before expecting an answer. Encourage them to tell you what they are thinking all the time. Then you will find out where and how they are going wrong. Remember, tutors also need time to think! If you are not sure, say so. You are not supposed to know everything.
(6) Question – ask helpful and intelligent questions which give clues, to stimulate and guide tutee thinking, and challenge their misconceptions. Do not say 'that's wrong!' – ask a question to give a clue. Ask 'why?' Examples of questions are: 'what kind of problem is this?', 'what are we trying to find out here?', 'can you state the problem in different words or a different way?', 'what important information do we already have?', 'can we break the problem into parts or steps?', 'how did you arrive at that?', 'does that make sense?', 'where was the last place you knew you were right?', 'where do you think you might have gone wrong?' and 'what kind of mistake do you think you might have made?'. Try to avoid: closed questions which require only a yes or no answer; questions which just rely on memory; questions which contain the answer; the question 'did you understand that?'; answering your own questions; indicating the 'difficulty' of any step.
(7) Make It Real – try to make the problem seem real and related to the life of your tutee. Ask the tutee to try to imagine what the problem would look like in real life. Some ideas to try: encourage tutees to use fingers, counters, cubes, sticks or any other objects to show the reality of the problem; have them draw dots, a picture, a list, table, diagram, graph or map; use charts including a number line, a multiplication matrix, and a place value chart; with your tutee's permission, mark their written working out with lines, arrows, colours, or numbering to help them; have your tutee think of what they have learned before or problems they have solved before relevant to the current problem; work through a similar but simpler problem; how can this problem be related to people, places, events and experiences in the home/community life of the tutee? (or those of someone they know or have seen on television); make up a similar problem using the pupil's own name and try to use everyday language.
(8) Summarise and Generalise – have your tutee summarise the key strategies and steps involved in solving the problem. Point out any errors or gaps, then summarise the key strategies yourself. Then talk about how these might be applied to another similar problem (generalised).

Previous research on Duolog Math

The math tutoring technique used here (Duolog Math) is relatively new and there is only modest literature on its effectiveness. It has proved considerably more difficult to develop a math tutoring procedure not purely focused on 'drills and skills' than has been the case with reading. The Duolog Math technique promotes dialogue between the pair and seeks to restrain the tutor from simply providing the right answer immediately. In this it has commonalities with tutoring techniques in reading, but some of its later (and more difficult) steps are very particular to Duolog Math.

Topping (2005) reported on questionnaires from 30 teachers who had engaged in cross-age peer-tutored Duolog Math projects with tutors aged 10–12 and tutees aged 7–8 years. However, teachers used math games for the first four weeks, before moving on to Duolog Math for the last four weeks. The majority of children showed improved co-operation with partner (86% of teachers), increased discussion of tasks (68%), increased enthusiasm and interest (55%), and increased concentration span (50%). A study is reported (Topping, Campbell, Douglas, & Smith, 2003) in which experimental tutees were successfully tutored in mathematical problem solving at home by their parent(s), while control children (n = 13) received traditional math problem homework, but this seems of doubtful relevance to peer tutoring. Thus, the previous Duolog Math reports are of small, short-term projects with self-selected participants, lacking implementation integrity data. In peer-tutored projects two different tutoring techniques were confounded. The present study dealt with these issues, by specifying technique, considering only peer tutoring, randomly allocating conditions, following children for two years and assessing quality of implementation.

Research questions

Three research questions were explored:

(1) With what level of integrity was the peer tutoring programme implemented (given that the pre-service training was modest in quantity and largely instructor-directed)?
(2) Was implementation integrity greater or less between Year 1 and Year 2 (the latter in largely different classes with largely different teachers)?
(3) To what extent were general classroom context indicators associated with the quality of implementation of individual technique, both within the observational schedule?

The last question stemmed from a desire to explore whether the individual technique was best learned directly from specific training by the teacher, or indirectly by percolation from the overall classroom ethos with a good deal of pupil initiative and peer observation in the development. The implications for the intensity of the initial training of teachers were then considered.

Method

The present study investigated a math tutoring technique (Duolog Math), a structured procedure but designed to be applicable to any mathematics curriculum material available in a school, but particularly mathematics problems. The tutoring

technique emphasised questioning, appropriate giving of positive feedback and correction of errors. Direct explaining was not emphasised owing to the uncertainty about tutor accuracy. The project trained teachers in this tutoring technique, but this training was very brief and didactic, since the sustainability of the project was important and releasing hundreds of teachers for extensive training was considered impractical. Teachers were, however, provided with a detailed resource pack. Then teachers were asked to train their pupils to implement their assigned condition of tutoring using modelling, practice and feedback.

Research design and sampling

The background project took place within a school district in Scotland of mixed socio-economic status with 143 elementary schools. It included both urban and rural areas, and thus schools were very various in size. Ethnic minority pupils were rare. The randomly allocated reading and mathematics interventions were offered to schools and a majority ($n = 125$) agreed to participate (several of the non-participating schools were changing head teacher or buildings) – thus a large self-selected sample that was then randomly allocated to condition. Random allocation was achieved with a random number generator. There was a control group for the background project but not for the investigation reported in this study.

The present paper is concerned only with mathematics in the mathematics and mathematics + reading conditions, which involved 80 schools. Thus 120 classes were involved (2×40 cross-age and 1×40 same-age) involving 3574 participants. The intervention involved either 7–8 or 10–11 year old pupils or both, so the whole classes involved were also pre-determined. In small schools where a single class contained pupils of two or more year groups, pupils not in the targeted ages were included in the intervention or not as the teacher decided, but further data on them were not gathered. Teachers involved represented the full range of age and experience. In some cases teachers participated reluctantly, having been 'volunteered' by their head teacher.

Intervention was in mathematics or mathematics + reading. Tutoring was either cross-age or same-age. Intervention was light (once per week) or intensive (three times per week), of about 20 minutes on task per session. Thus, about 10 schools were in each of the eight conditions, totalling 80 schools. Some children moved schools and were excluded ($n = 188$, 5.3%), but these were from a great many schools. A few schools dropped out and some changed condition and were excluded (a total of 10 schools), but as these schools were scattered across conditions, data gathering was not greatly affected and the influence on the power of the study was minimal. The intervention was for 15 weeks (January–April), repeated in the second year for the same pupils, many of whom had a new class and new teacher.

For the present study, in Year 1 observation of implementation integrity occurred in 29 randomly allocated schools (out of the 80 participating). In Year 2, observations occurred in a (different) randomly allocated 30 schools (out of the 80 participating). Observation was thus conducted in 59 classes involving a total of 1770 pupils. Performance feedback was not given to teachers on the basis of these observations. Data gathering occurred mid-intervention in February in both years. All observations were completed within a three-week period.

Training

Participating teachers received half a day of continuing professional development on Duolog Math. At least two teachers were invited from every school; sometimes this was the intervention teacher and the head teacher. Training involved a context-setting talk from a senior manager of the school district, an outline of the rationale and importance of randomised controlled trials, a talk about the tutoring technique and how to organise it, and an opportunity to discuss with other schools in the same condition. It was largely didactic. A pack of detailed materials was provided to each school for further study. These materials referred briefly to operational definitions and the research background, but were mainly practical materials for teachers to give to pupils and organisational advice for the teacher her/himself. Subsequently, a research assistant visited schools on request to offer support and saw other teachers in group 'surgery' sessions. After the intervention, teachers received a further half day of continuing professional development, involving presentations from teacher colleagues who had implemented successfully and further opportunity to discuss. This was repeated in the second year for new teachers, adding use of a demonstration video and brief presentations from two selected first year schools to initial training.

Observation measures

The variables observed were either class-wide contextual measures of teacher–pupil interaction (29 variables) or quality of individual pair technique (11 variables), measured by direct researcher observation in each class and manual recording onto a schedule.

Contextual variables were divided into 'General', 'Resources', 'Introduction to the Session', 'During the Session' and 'End of Session' (see Table 1). General and End of Session variables thus mostly involved observation of teacher and pupils, Resources and Introduction involved mostly observation of the teacher, and During the Session mostly observation of pupils. These items were recorded on a categorical scale 1 to 4: not at all true of this class/lesson, partly true of this class/lesson, mostly true of this class/lesson, completely true of this class/lesson. Thus, the expected result for each variable would be approximately three. The variables observed related directly to aspects of organisation that had formed part of the continuous professional development for teachers, and as such had face validity. However, no independent large-scale measures of reliability or validity were available.

Observation of the quality of technique of the individual pair was focused on 11 variables which were recorded almost entirely as frequency counts (see Table 2). The expected result would be: about two for Reading, Checking, Pause for Think-Aloud and Making It Real, about four to six for Listening, Questioning and Praise, and about one and a half for Summarising and Generalising. These factors were aspects of the technique and as such had high face validity.

Given that this was the first study of implementation of this tutoring technique and it was not clear what might be of value, a wide range of variables were included and observations were based on frequencies rather than durations. It was considered that a further study might focus on a smaller number of variables found to be of especial interest and use some measures of duration.

Table 1. First and second year observations: whole class.

Observed variable	First year (n = 29) Mean	*SD*	Second year (n = 30) Mean	*SD*	Effect size
General					
Do seating positions enable tutors/tutees to work together?	3.97	0.19	4.00	0.00	−0.16
Ethos of the classroom: warm and lively?	3.90	0.41	3.63	0.67	0.66
Is the venue reasonably quiet?	3.62	0.78	3.50	0.90	0.15
Were the pairings decided according to ability?	3.50	0.84	3.39	0.58	0.13
Are the terms 'tutor' and 'tutee' used?	3.25	1.32	3.52	1.12	−0.20
Do pairs have a Duolog Math folder/jotter to record work in?	3.22	1.16	2.70	1.21	0.45
Resources					
Are a variety of problems available?	2.24	1.43	2.10	1.27	0.10
Is the tutee choosing the problems?	1.75	1.21	1.65	1.23	0.07
Are problems leveled?	2.00	1.25	2.00	1.12	0.00
Does the teacher introduce the problem?	1.59	1.15	2.27	1.41	−0.59
Is there discussion: things to remember during Duolog Math?	2.14	1.36	1.87	1.14	0.20
Does the teacher suggest/provide questions/ways to make it real?	1.48	1.06	1.47	1.04	0.01
Does the teacher suggest/provide any materials?	1.31	0.93	1.80	1.35	−0.53
During the session					
Do the pairs appear to be mainly on task?	3.76	0.51	3.23	0.90	1.04
Prompt cards/posters to remind pupils of strategies?	2.76	1.38	2.40	1.40	0.26
Do the children appear to refer to these?	1.50	0.62	1.29	0.77	0.34
Is there movement to gather materials for 'Make it Real'?	1.93	1.07	1.90	1.06	0.03
Are tutors asking for the answer sheet to 'Check'?	2.11	1.26	2.00	1.31	0.09
Is the teacher monitoring all pairs throughout the session?	3.34	0.94	3.47	0.90	−0.14
Are the pupils doing 1 problem per session?	2.76	1.38	1.90	1.21	0.62
Do any pairs stop working on the problem early?	2.67	1.21	2.93	1.36	−0.21

(*Continued on next page*)

Table 1. (*Continued*)

Observed variable	First year (n = 29)		Second year (n = 30)		Effect size
	Mean	*SD*	Mean	*SD*	
End of session					
Is there evidence of Summarising/Generalising?	1.45	0.83	1.70	0.95	−0.30
Are the diary records used?	2.70	1.49	2.30	1.44	0.27
Are the comments it contains positive?	3.25	0.75	3.08	1.08	0.23
Including comments from the teacher?	1.62	1.12	1.64	1.28	−0.02
Is there a plenary session?	1.82	1.34	2.33	1.42	−0.38
Do pairs participate in this?	4.00	0.00	3.71	0.85	
Do pairs keep a record of answer/summary of steps taken?	2.07	0.62	2.10	0.71	−0.05

Table 2. First and second year observations: individual pairs.

	First year (n = 29)			Second year (n = 30)			
	Cases where present	Mean	*SD*	Cases where present	Mean	*SD*	Effect size
Length of session (in minutes)	27	21.07	6.77	30	21.80	8.50	−0.11
No. of problems completed during session	26	1.35	0.56	28	1.89	1.45	−0.96
Number of instances of Read	22	1.95	1.05	23	2.43	1.47	−0.46
Number of instances of Listen	0	0.00	0.00	2	1.00	0.00	
Number of instances of Check	12	1.75	1.60	14	1.43	0.65	0.20
Number of instances of Praise	10	2.20	1.69	9	2.33	1.87	−0.08
Number of instances of Pause for Think-Aloud	18	1.78	0.94	14	3.36	2.90	−1.58
Number of instances of question	24	4.21	2.40	25	6.08	5.59	−0.78
Number of instances of Make it Real	24	1.67	1.01	23	2.43	2.59	−0.75
Number of instances of Summarise	6	1.17	0.41	8	1.25	0.46	−0.20
Number of instances of Generalise	1	1.00	0.00	5	1.00	0.00	

Procedure – intervention

In cross-age tutoring between two classes of different ages, pupils in each class were ranked by math ability, the most able tutor in the older class matched with the most able tutee in the younger class, and so on. In same-age tutoring, one class was ranked by math ability, divided into tutors above and tutees below, and the most able tutor matched with the most able tutee, and so on. This meant that in same-age classes the weakest tutee was helped by an average tutor, while in cross-age classes the weakest tutee was helped by the weakest older tutor. Small matching adjustments were made on grounds of social compatibility. Parents were informed of the project by letter, and further discussions were held with a very small minority of parents. No parents withdrew their children.

Teachers then trained participating pupils together in the technique and began the project. Teachers were asked to tell pupils about the structure of the technique and give a demonstration of how to do it (in Year 1 by role-playing adults, but by Year 2 a training video was available). Pairs then immediately practised the technique with math problems of appropriate difficulty, while teachers circulated to monitor, coach and give feedback. During subsequent intervention sessions, absence of pupils for any session required teachers to rematch un-partnered individuals. Teachers continued monitoring and support (a teacher checklist for this was provided). Pairs kept a brief written record of their math work. After 15 weeks, children were gathered together and asked if and how they wished to continue.

Procedure – research

Observation of implementation integrity was undertaken in both years. The same researcher made all observations. She made one visit to each class in the middle of the intervention period. Recording was done manually on a structured observation schedule. Each observation continued for the whole of the Duolog Math session (in Year 1 averaging 21 minutes and in Year 2 almost 22 minutes) and immediately after it. The General and Resources sections of whole class contextual observations were completed in the first five minutes of the tutoring session. The researcher then observed the whole class and teacher in interaction (for an estimation of general classroom ethos) and also observed one tutoring pair (for purity of technique). Thus episodes of whole class contextual during observation alternated with episodes of individual observation of the targeted pair (the teacher divided all the pairs into top, middle and bottom thirds of ability and the researcher randomly chose one pair to observe per class). The whole class contextual end of session section was completed after the session had ended.

Analysis

Contextual results were aggregated at the level of the class, while individual results were at the level of a pair of pupils. Descriptive statistics led to comparisons between years using paired sample t tests (the random samples being treated as equivalent). Regression analysis investigated whether contextual variables related to quality of technique in Individual pairs.

Results

Whole class contextual observations

In general, the mean whole class contextual observations for Year 1 and Year 2 were similar, only 5 of 29 variables showing significant differences between Years 1 and 2 (Table 1). However, in three of these cases the differences between Year 1 and Year 2 were negative. These were: Does the teacher introduce the problem $t = -2.287$, $n = 16$, $p = 0.036$; and Does the teacher suggest ways to make it real $t = -2.400$, $n = 16$, $p = 0.027$. Another was negative but indicated Year 2 pupils did better: Do pupils stop working on problems early $t = -2.426$, $n = 16$, $p = 0.027$. Another was positive and indicated Year 2 pupils did better: Do pupils do one problem per session $t = 2.263$, $n = 16$, $p = 0.038$. The last of these (Comments in the record positive) was based upon very small numbers. Overall then, there was little difference between years. This was corroborated by inspection of effect sizes, which were mostly small but if anything favoured the Year 1 teachers.

In both years, seating positions, ethos and quietness were observed to be high. High rates of time on task were evident. Teachers were generally good at monitoring all pairs. Pairings were generally decided according to ability. Mostly, the terms 'tutor' and 'tutee' were used. Pairs had a jotter to record their work in. Dairies were somewhat used, but they contained positive comments in a minority of cases. Plenary sessions were somewhat used, but the observed pair participated in these in a minority of cases, especially in Year 1.

However, a very modest variety of problems was available and problems were not necessarily levelled (categorised according to their difficulty). Tutees chose the problems in a minority of cases. Teacher introduction of problems was poor, especially in Year 1. Reminders of 'Things to Remember' were not very frequent, and teachers rarely suggested ways to 'Make it Real' or provided any materials. A majority of teachers had prompt cards/posters available, but a minority of children were observed to refer to these. A majority of children were only doing one problem per session in Year 1, and about half of the children in Year 2 (this was principally associated with the teacher not providing sufficient problems for the class). A majority of children in both years stopped work early. Crucially, there was little evidence of Summarising or Generalising.

Individual pair observations

The mean individual pair technique practice quality observations for Year 1 and Year 2 were similar. None showed a significant difference. However, in all but one case Year 2 was higher than Year 1 (Table 2, Figure 1), only one effect size was positive, and some negative effect sizes were quite substantial. Of course, this might be because the pupils were more mature in Year 2 (i.e. an age effect), but this would be more plausible with regular teaching rather than a special intervention. Questioning was high in both years but higher in Year 2. Reading the Question and Making it Real were about central on the four-point scale. Most children only completed one problem per session. Praising occurred fairly frequently, but only for a small number of children. Pausing for Think-Aloud was much higher in Year 2 than Year 1, but again not all children were involved. Checking occurred in few cases and then at moderate frequency. Summarising and Generalising were recorded for

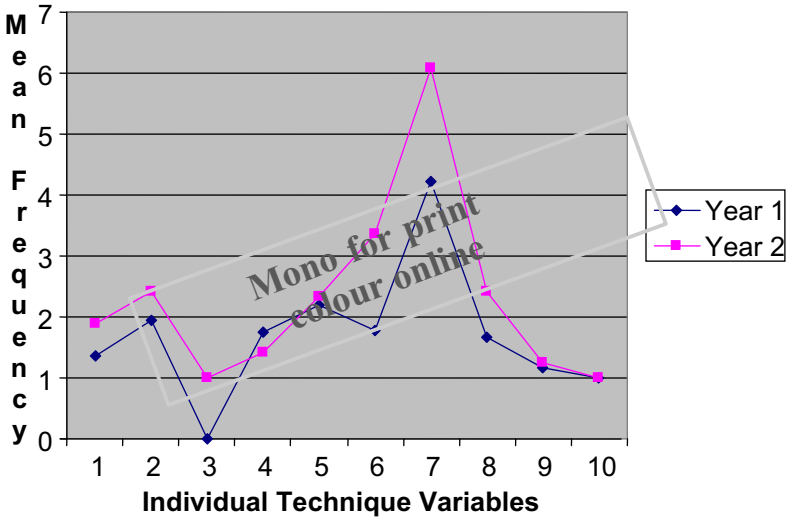

Figure 1. Individual technique variables in Years 1 and 2.

only a small number of children. Listening was almost never recorded, but this was a very difficult behaviour to observe.

Relationship between contextual and individual observations

Did teachers with better classroom organisation obtain better performance from individual pairs? In order to explore whether whole class contextual variables had any effect on the quality of practice of individual pairs, for each pupil their frequency counts (fc) on the nine individual pair technique variables which related to quality of practice (excluding the first 2 of the 11) were summed (0–fc × 9) (averaging them would have been equivalent), yielding an overall index of pupil Practice Quality. Of course, frequencies varied between variables, so some variables contributed more than others. Each contextual variable for both Years 1 and 2 was then compared to this overall index of Practice Quality using linear regression. In Year 1 three contextual variables were found to be associated with higher Practice Quality (all $df = 1.27$): Are comments the diary contains positive? ($\beta = 0.585$, $R^2 = 0.342$, $F = 5.207$, $p = 0.046$), Is there evidence of Summarising and Generalising? ($\beta = 0.454$, $R^2 = 0.206$, $F = 6.992$, $p = 0.013$), and Is there movement around the classroom to gather materials? ($\beta = 0.394$, $R^2 = 0.155$, $F = 4.947$, $p = 0.035$). Are the terms tutor and tutee used? almost reached statistical significance ($\beta = 0.337$, $R^2 = 0.114$, $F = 3.341$, $p = 0.079$). Are prompt cards and posters available to remind pupils of strategies? proved negative ($\beta = -0.415$, $R^2 = 0.415$, $F = 5.629$, $p = 0.025$).

In Year 2, however, a completely different set of variables emerged (all $df = 1.27$). Do the pairs appear mainly on task? ($\beta = 0.443$, $R^2 = 0.197$, $F = 6.606$, $p = 0.016$) and Number of problems completed during the session ($\beta = 0.411$, $R^2 = 0.169$, $F = 5.074$, $p = 0.033$) were positively related to Practice Quality. Are pupils doing one problem per session? was negatively related to Practice Quality ($\beta = -0.437$, $R^2 = 0.191$, $F = 6.356$, $p = 0.018$), as would be expected. Two further variables were negatively related: Is there a discussion of things to remember? ($\beta = -0.456$, $R^2 = 0.208$,

$F = 7.103$, $p = 0.013$) and Does the teacher introduce the problem? ($\beta = -0.438$, $R^2 = 0.192$, $F = 6.415$, $p = 0.017$). It is difficult to see why this should be so.

Discussion

Teachers and pupils in both years managed to implement some but not all of the technique effectively. It is possible that the motivation of the participating children was declining in their second year of involvement (or the random selection of different children for observation might have introduced some bias). However, there was little evidence for this in the observational data. Indeed, the evidence suggested that while Year 2 teachers were very similar in behaviour to Year 1 teachers, the pupils in Year 2 were somewhat better at implementing the technique (with which of course they had been familiar in Year 1). More problems were addressed in Year 2. This might have been the result of greater familiarity in Year 2 teachers. Although they received no more continuous professional development than Year 1 teachers, they had colleagues experienced in the method within the school to support them. To what extent this facilitated the transfer of technical information and to what extent it facilitated the transfer of a sense of reassurance that the process was not new is an interesting question – perhaps for future research. The individual technique variables in five cases aligned with the expectations noted prior to the study, but Listening was grossly under-recorded, Praise was less than expected, and Summarisation and Generalisation were lower than expected. These findings may be considered generalisable to non-ethnic-minority pupils aged 7–8 and 10–11 in elementary schools of all sizes elsewhere in the country.

Limitations

The unavailability of non-experimental schools using the technique did not permit the extensive piloting of the observation schedule, although one part of it was based on aspects of classroom organisation in which teachers were trained and the other mirrored exactly the elements of the individual technique, so both had high face validity. Reliability and validity of the observation scales was not established, and this is necessary in the future. The consistency of the single researcher conducting the observations is not known, which could have been addressed through involvement of a second observer in a proportion of classrooms (albeit at a cost). It would have been possible to conduct *post hoc* reliability analyses if the observed pupils had been the same in Year 1 and Year 2, but this was not the case. Similarly, it would have been possible to analyse for systematic observer drift across the two years, a potential explanation for some of the differences in pupil behaviours observed across the two years. Certainly, comparing random samples (but not the same pupils) in Years 1 and 2 potentially introduced other confounding variables into the analysis.

Relationship to literature

Relating these findings back to the literature is informative. A high degree of instructional standardisation was attempted by coupling the training with a detailed resource pack (Rowan & Miller, 2007). Training attempted to establish teams within schools (Blom-Hoffman, 2007), but might not always have been successful. Key features of

individual technique were emphasised (Pankratz et al., 2006) – nine points to remember with some harder than others. Researchers worked with teachers on implementation, but not with all teachers and largely on the teacher's definition of the problem. There was no performance feedback based on direct observation, either short-term or persistent (Duhon et al., 2009). Brief didactic methods were used in training the teachers, even though less effective than lengthier (and costlier) modelling, practice and feedback. However, the teachers were to train the pupils by the latter method. Partial implementation is common and this study was no different (Taylor & Teddlie, 1999). Teachers tended to implement the elements of the technique that were less challenging faster and under-emphasised the more complex instructional processes of Summarising and Generalising – of course this is not surprising.

Future research

Future research should explore variations in the length and type of training to ascertain relative influence on implementation integrity. Variables with surprising negative associations should be further explored. Comparisons of cost-effectiveness of different intensities of training would become possible. Key elements of procedure should be emphasised. Collecting implementation integrity data on all participants would be advantageous. There is a case for collecting such data on more than one occasion. Reliability and validity of the observation scales and inter-rater reliability should be established. Replication elsewhere on this scale would be advantageous. Collecting outcome data and relating it to implementation integrity data would be valuable. In terms of analysis, hierarchical linear modelling would be useful to explore effects for pupils nested in classrooms nested in conditions, but large numbers would be necessary.

Practice implications

Tutoring in this form can be implemented as a result of very brief training for teachers, but less than perfectly. Training for teachers involving modelling, practice and feedback should be considered, although this would take more time and cost. Further training should focus on those parts which teachers appeared to find hardest to implement. Many problems should be available and problems should be levelled. Teacher behaviours in Introducing the problem, suggesting ways to Make it Real, and holding Plenary Sessions should be focused upon. Summarising and Generalising especially need further emphasis. Teachers should mention things to remember and include their comments in diary records. Pupils should complete more than one problem per session and not complete before time is up. Listening, Praising and Checking require more consistent implementation. Organisational options (same-age or cross-age; light or intensive) are available for teachers constrained by their circumstances. The technique may work in ways other than merely increasing the amount of practice in math, including inducing social and emotional changes in the children. These variables are being studied in further research on this project.

Conclusions

Year 1 and Year 2 implementation observations were somewhat similar, indicating partial implementation in both years, although somewhat better for pupils in Year 2.

Seating positions, ethos, quietness, high time on task, teacher monitoring of pairs and pairings decided according to ability were relatively high. However, problem availability was difficult and problems were not likely to be levelled. There were deficits in: teachers introducing the problem, suggesting ways to make it real, providing additional other problems and holding plenary sessions. Crucially, there was very little Summarising or Generalising. Many pupils completed only one problem per session, often before time was up.

Considering to what extent contextual variables could predict individual variables, those positively related included: Are comments the diary contains positive, Is there movement around the classroom to gather materials, Do the pairs appear mainly on task and Number of problems completed during the session. Variables negatively related (as expected) included: Are pupils doing one problem per session. These should be borne in mind by teachers seeking to determine to what extent classroom variables can compensate for specific training.

Thus, with regard to the research questions:

(1) The tutoring technique was only partly implemented. Some aspects were done well, others poorly. Thus, the brief and largely instructor-directed training for teachers appeared to have partial effects.
(2) The peer tutoring programme was implemented with similar integrity in Year 1 and Year 2, but Year 2 was somewhat better for pupil implementation (with largely different teachers).
(3) Some General classroom context indicators were associated with quality of implementation of individual tutoring technique behaviours, both within the observational schedule. However, some were negative as well as positive. It must be concluded that this association was weak and mixed.

There are implications here for the quantity and quality of training. It seems likely that two half days of pre-training for teachers including modelling, practice and feedback would have been more effective than the single half day of pre-training offered, without making the intervention so time-intensive for teachers that it became cost-ineffective.

References

Arkoosh, M.K., Derby, K.M., Wacker, D.P., Berg, W., McLaughlin, T.F., & Barretto, A. (2007). A descriptive evaluation of long-term treatment integrity. *Behavior Modification, 31*, 880–895.

Auld, R.G., Belfiore, P.J., & Scheeler, M.C. (2010). Increasing pre-service teachers' use of differential reinforcement: Effects of performance feedback on consequences for student behavior. *Journal of Behavioral Education, 19*(2), 169–183.

Blom-Hoffman, J. (2007). School-based promotion of fruit and vegetable consumption in multiculturally diverse, urban schools. *Psychology in the Schools, 45*(1), 16–27.

Bradshaw, C.P., Reinke, W.M., Brown, L.D., Bevans, K.B., & Leaf, P.J. (2008). Implementation of school-wide positive behavioral interventions and supports (PBIS) in elementary schools: Observations from a randomized trial. *Education and Treatment of Children, 31*(1), 1–26.

Britz, M.W., Dixon, J., & McLaughlin, T.F. (1989). The effects of peer tutoring on mathematics performance. A recent review. *B. C. Journal of Special Education, 13*(1), 17–33.

Burns, M.K., Peters, R., & Noell, G.H. (2008). Using performance feedback to enhance implementation fidelity of the problem-solving team process. *Journal of School Psychology, 46*(5), 537–550.

Cohen, P.A., Kulik, J.A., & Kulik, C.-L.C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal, 19*(2), 237–248.

Digennaro, F.D., Martens, B.K., & Kleinmann, A.E. (2007). A comparison of performance feedback procedures on teachers' treatment implementation integrity and students' inappropriate behavior in special education classrooms. *Journal of Applied Behavior Analysis, 40*(3), 447–461.

Duhon, G.J., Mesmer, E.M., Gregerson, L., & Witt, J.C. (2009). Effects of public feedback during RTI team meetings on teacher implementation integrity and student academic performance. *Journal of School Psychology, 47*(1), 19–37.

Gansle, K.A., & McMahon, C.M. (1997). Component integrity of teacher intervention management behavior using a student self-monitoring treatment: An experimental analysis. *Journal of Behavioral Education, 7*, 405–419.

Gilbertson, D., Witt, J.C., Singletary, L.L., & VanDerHeyden, A. (2007). Supporting teacher use of interventions: Effects of response dependent performance feedback on teacher implementation of a math intervention. *Journal of Behavioral Education, 16*(4), 311–326.

Ginsburg-Block, M.D., Rohrbeck, C.A., & Fantuzzo, J.W. (2006). A meta-analytic review of social, self-concept, and behavioral outcomes of peer-assisted learning. *Journal of Educational Psychology, 98*(4), 732–749.

Gresham, F.M. (1989). Assessment of treatment integrity in school consultation and prereferral intervention. *School Psychology Review, 18*, 37–50.

Noell, G.H., Witt, J.C., Slider, N.J., Connell, J.E., Gatti, S.L., Williams, K.L., … Resetar, J.L. (2005). Treatment implantation following behavioral consultation in schools: A comparison of three follow-up strategies. *School Psychology Review, 34*, 87–106.

Pankratz, M.M., Jackson-Newsom, J., Giles, S.M., Ringwalt, C.L., Bliss, K., & Bell, M.L. (2006). Implementation fidelity in a teacher-led alcohol use prevention curriculum. *Journal of Drug Education, 36*(4), 317–333.

Pence, K.L., Justice, L.M., & Wiggins, A.K. (2008). Preschool teachers' fidelity in implementing a comprehensive language-rich curriculum. *Language, Speech, and Hearing Services in Schools, 39*, 329–341.

Person, N.K., & Graesser, A.G. (1999). Evolution of discourse during cross-age tutoring. In A.M. O'Donnell & A. King (Eds.), *Cognitive perspectives on peer learning* (pp. 69–86). Mahwah, NJ: Lawrence Erlbaum.

Robinson, D.R., Schofield, J.W., & Steers-Wentzell, K.L. (2005). Peer and cross-age tutoring in math: Outcomes and their design implications. *Educational Psychology Review, 17*(4), 327–362.

Rohrbeck, C.A., Ginsburg-Block, M.D., Fantuzzo, J.W., & Miller, T.R. (2003). Peer-assisted learning interventions with elementary school students: A meta-analytic review. *Journal of Educational Psychology, 95*(2), 240–257.

Roscoe, R.D., & Chi, M.T.H. (2007). Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions. *Review of Educational Research, 77*(4), 534–574.

Rowan, B., & Miller, R.J. (2007). Organizational strategies for promoting instructional change: Implementation dynamics in schools working with comprehensive school reform providers. *American Educational Research Journal, 44*(2), 252–297.

Sharpley, A.M., & Sharpley, C.F. (1981). Peer tutoring: A review of the literature. *Collected Original Resources in Education, 5*(3), C-11.

Stead, M., Stradling, R., Macneil, M., Mackintosh, A.M., & Minty, S. (2007). Implementation evaluation of the Blueprint multi-component drug prevention programme: Fidelity of school component delivery. *Drug and Alcohol Review, 26*(6), 653–664.

Sterling-Turner, H.E., Watson, T.S., Wildmon, M., Watkins, C., & Little, E. (2001). Investigating the relationship between training type and treatment integrity. *School Psychology Quarterly, 16*(1), 56–67.

Taylor, D.L., & Teddlie, C. (1999). Implementation fidelity in Title I schoolwide programs. *Journal of Education for Students Placed at Risk (JESPAR), 4*(3), 299–319.

Topping, K.J. (2005). *Problem-solving evaluation*. Retrieved March 10, 2009, from http://www.dundee.ac.uk/eswce/research/projects/problem-solving/evaluation.

Topping, K.J., & Ehly, S. (Eds.). (1998). *Peer-assisted learning*. Mahwah, NJ: Lawrence Erlbaum.

Topping, K.J., Campbell, J., Douglas, W., & Smith, A.J. (2003). Cross-age peer tutoring in mathematics with 7 & 11 year olds: Influence on mathematical vocabulary, strategic dialogue and self-concept. *Educational Research, 45*(3), 287–308.