Grünke, M., Boon, R. T., & Burke, M. D. (2015). Use of the randomization test in single-case research. *International Journal for Research in Learning Disabilities, 2*(2), 44-64.

Use of the Randomization Test in Single-Case Research

Matthias Grünke
University of Cologne, Germany

Richard T. Boon
The University of Texas at San Antonio

Mack D. Burke
Texas A&M University

Abstract

The purpose of this study was to illustrate the use of the randomization test for single-case research designs (SCR; Kratochwill & Levin, 2010). To demonstrate the application of this approach, a systematic replication of Grünke, Wilbert, and Calder Stegemann (2013) was conducted to evaluate the effects of a story map to improve the reading comprehension skills of five elementary students with learning disabilities in Germany. A multiple-baseline design (Baer, Wolf, & Risley, 1968) was used to evaluate the effectiveness of the story mapping instruction in teaching students the key story grammar elements in a reading passage. However, unlike in traditional multiple-baseline designs, intervention and withdrawal phases were applied at randomly determined points. Results indicated that the use of story maps increased the students' recall and comprehension of the stories from baseline to intervention, and continued during maintenance. A randomization test confirmed that the differences between baseline and intervention were statistically significant. Findings, limitations, and implications of the use of randomization tests in SCR are discussed.

The field of special education is moving toward identifying and implementing evidence-based practices (EBPs) in the classroom (e.g., Cook, Tankersley, Cook, & Landrum, 2008; Cook, Tankersley, & Harjusola-Webb, 2008; Odom et al., 2005; Torres, Farley, & Cook, 2014). As a result, decisions regarding the selection of instructional strategies to meet the diverse

needs of students with disabilities are increasingly being supported by research from the special education literature (Cook & Cook, 2013; Cook, Tankersley, & Landrum, 2013). One particular area of research is the importance of establishing quality standards that determine what constitutes an EBP (Cook et al., 2014; Cook, Tankersley, & Landrum, 2009). In order to be counted as an EBP, an approach must have documented effectiveness. Specifically, it has been suggested that to qualify as an EBP, an intervention must use high-quality research designs consisting of at least two group design experiments (Gersten et al., 2005) or a series of at least five single-case studies (Horner et al., 2005).

With regard to students with learning disabilities (LD), single-case research (SCR) is playing an increased role in establishing the evidence base for practices. SCR uses the data from one participant or from a very small number of subjects to establish the existence of cause-and-effect relationships (Riley-Tillman & Burns, 2009). In other words, SCR is used to determine whether there is a functional relationship between manipulation of the independent variable(s) and corresponding changes in the dependent measure(s). Further, in SCR designs, individuals serve as their own controls by providing a baseline prior to implementation of an approach. Thus, an intervention can be progress monitored and responsiveness to intervention can be judged against prior progress. As a result, researchers are able to make inferences regarding whether a given intervention is instrumental in fostering different skills for a specific student (Riley-Tillman & Burns, 2009).

Several meta-analyses in recent years have focused on interventions for students with LD that are mainly based on SCR (e.g., Codding, Burns, & Lukito, 2011; Lee & Kim, 2013; Zheng, Flynn, & Swanson, 2013). One reason why this research methodology seems to be especially applicable for evaluating interventions for students with LD is that repeated measurement of respective success criteria (e.g., reading fluency, automatic recall of math facts, spelling skills) is often relatively easy. To frequently record school performance-related variables like reading fluency, automatic recall of math facts, or spelling skills in an objective, reliable, and valid way is generally a lot less complicated than to register emotional or social parameters (self-esteem, interpersonal skills, or psychological resilience). Through systematic measurement of the dependent measure, intervention responsiveness can be evaluated.

However, despite the prominent role that SCR occupies in the research-based literature for students with LD, many scholars still have reservations about this approach (Matson, Turygina, Beighleya, & Matson, 2012). A major reason for such reservation involves the common method of analyzing data from SCR, which relies on visual analysis. Visual analysis consists of graphing a given data set and then appraising the differences between phases for changes in level, trend, and variability. While widespread in the SCR literature, visual analysis of SCR data is viewed by some researchers as being biased (Dugard, File, & Todman, 2012).

The inter-rater reliability for this approach is alarmingly low, rarely exceeding .50 (Brossart, Parker, Olson, & Mahadevan, 2006). Further, the effect of training and experience on visual inspection seems to be negligible. For example, Harbst, Ottenbacher, and Harris (1991) demonstrated that long-time journal reviewers performed little better than completely untrained raters at graph judgment tasks.

For these reasons, several attempts have been made to apply statistical inferential procedures to analyzing data from SCR designs (Callahan & Barisa, 2005; Campbell & Herzinger, 2010; Ferron, 2002; Garthwaite & Crawford, 2004; Janosky, Al-Shboul, & Pellitieri, 1995; Manolov, Arnau, Solanas, & Bono, 2010; Parker, Vannest, Davis, & Sauber, 2011; Scruggs, Mastropieri, & Regan, 2006). Unfortunately, applying common parametric tests for group comparisons is generally unsuitable in this context (Campbell & Herzinger, 2010), primarily due to the small sample size and the repeated measures that are used in SCR. Thus, we need to look for alternatives for common parametric methods to provide researchers with tools to analyze their data from SCR designs in a way that is equally objective.

Randomization Tests for SCR

One remedy for making visual analysis more acceptable is to apply randomization tests for phase designs (Kratochwill & Levin, 2010). Traditionally, in SCR, it is recommended that baseline observations continue until a stable pattern of measurements has been established (Gast & Ledford, 2010). However, when researchers wish statistically to analyze their data with a randomization test for phase designs, a different approach is needed – one that first introduces a random assignment scheme into the experiment (Ferron, Foster-Johnson, & Kromrey, 2003).

One way to introduce randomization in phase designs consists of defining a total number of measurement points, a reasonable minimum and maximum number of measurement points for each phase a priori, and subsequently selecting the beginning of the intervention phase by chance within the preset range of options (Edgington, 1992). Randomization tests work by taking into account all possible permutations of the data. For example, if a simple AB design with 30 observations and at least 5 measurement points in each phase was applied, the intervention could commence after the 5^{th}, 6^{th}, 7^{th}, 8^{th}, … or the 25^{th} baseline observation. This would add up to 21 possible intervention starting points. One of these options is drawn by chance, and the study is executed. Following the collection of all data, a test statistic is computed, and the results are compared to the test statistic of the 20 theoretically possible permutations that could have occurred. The test p-value is the proportion of test statistic values greater than or equal to the observed test statistic. In the current example, if the observed test statistic exceeded the other 20 options, the probability of such an outcome would be 1/21 = 0.048 (Todman, 2002).

Other designs and computational procedures are often much more complicated than this simple example. However, in every instance, researchers do not rely on a standard test statistic distribution (e.g., *t*- or *F*-distribution), but act on the premise that all the information needed to perform inferential statistics is included the data set itself, and derive a *p*-value from permutations of the data. In the aforementioned example, this can easily be done with a pocket calculator. However, such a procedure can otherwise become very complex. For instance, for an AB multiple-baseline design across five participants again with 21 possible intervention starting points per subject, there would be $21^5 = 4{,}084{,}101$ permutations. Enumerating such a large number of possibilities requires fast and intensive computing.

Until recently computational power to perform such randomization tests has not been readily available, which explains why these methods have not yet played a major role in analyzing data from SCR designs in practical research (Dugard et al., 2012). Nowadays, Monte Carlo randomized tests, which compute an approximate p-value of the observed test statistic for a random sample of all possible data arrangements, are commonly used. The test p-value is the proportion of test statistic values in the random sample as large as the observed test statistic. Monte Carlo randomized tests are available in the familiar environments of IBM® SPSS and Microsoft® Excel (Dugard, 2013), or SCDA, a software application for analyzing single-case designs implemented in R, which includes a randomization test module, SCRT (Bulté & Onghena, 2013).

Story Mapping

A story map is a visual strategy designed to promote comprehension of the main parts of a story. It has been shown to be an effective intervention for elementary students with LD (e.g., Boulineau, Fore, Hagan-Burke, & Burke, 2004; Gardill & Jitendra, 1999; Idol, 1987; Idol & Croll, 1987; Johnson, Graham, & Harris, 1997; Stagliano & Boon, 2009; Taylor, Alber, & Walker, 2002; Wade, Boon, & Spencer, 2010). According to Davis and McPherson (1989), a story map is "… a graphic representation of all or part of the elements of a tale and the relationships between them" (p. 232). It is a form of a graphic organizer that makes the structure of concepts and relationships between them apparent by creating a systematic schema to connect prior knowledge with the content of a text that a learner is reading (Anderson & Pearson, 1984; Ausubel, 1960, 1968).

Story maps reduce the amount of semantic information a student has to process in order to extract meaning (Jitendra & Gajria, 2011), thus decreasing the potential for cognitive overload (O'Donnell, Dansereau, & Hall, 2002). Figure 1 shows a sample story map template from Idol (1987, p. 199) completed with all the main components of the fairy tale *The Frog Prince* (Grimm & Grimm, 2013).

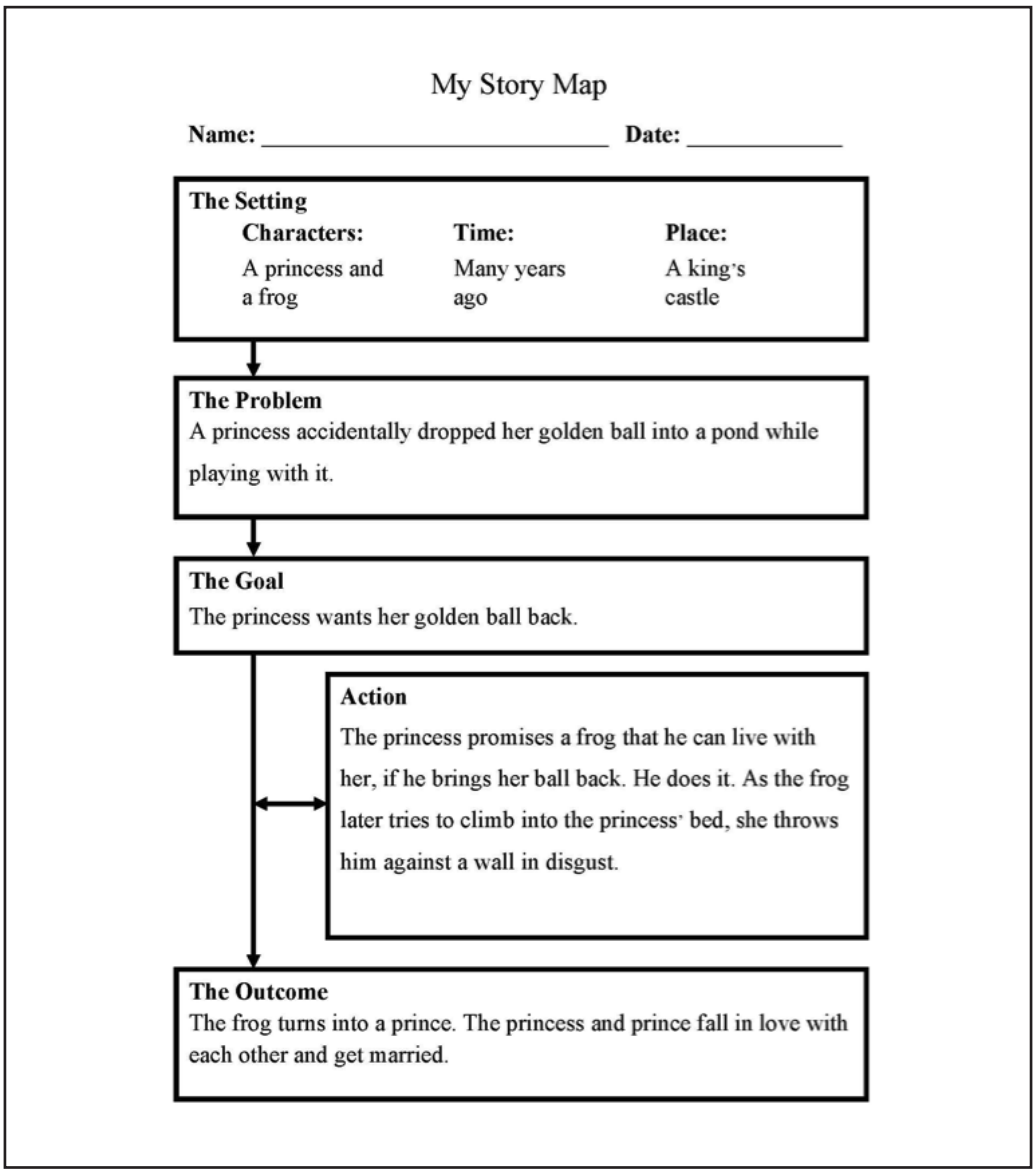

Figure 1. Sample story map from the story *The Frog Prince*. Adapted from "Group story mapping: Comprehension strategy for both skilled and unskilled readers", by L. Idol, 1987, *Journal of Learning Disabilities, 20*(4), 199. Copyright 1987 by SAGE Publications. Adapted with permission.

Purpose

The primary purpose of the present study was to illustrate the use of the randomization test for SCR designs (Dugard, 2013; Kratochwill & Levin, 2010). Since this approach has only been applied four times (Grünke & Calder Stegemann, 2014; Grünke et al., 2013; Mastropieri, Scruggs, Mills, et al., 2009; Regan, Mastropieri, & Scruggs, 2005), our study extends the current special education research base on its application to LD populations. A secondary purpose of the study was to conduct a systematic replication of a previous study by Grünke et

al. (2013) examining the effects of a story map to improve the reading comprehension skills of elementary students with LD, which also extends current research on the use of a randomization test with a story mapping intervention.

Method

Participants

Students. Five elementary students were recruited to participate in the study. All had been identified with an LD by a multidisciplinary team. In Germany, where the study took place, the main criterion for diagnosing a student with LD is generalized school failure. Such a label indicates that a student shows deficits in one or more psychological processes that manifest themselves in an insufficient ability to perform at grade level in most or all academic core courses compared to typically achieving peers. However, these students do not meet the criteria for a mild intellectual disability, even though often they exhibit low average intelligence (Al-Yagon, Cavendish, Cornoldi, et al., 2013).

Participants included two female (Asena and Julia) and three male students (Eman, Leon, and Marvin). Two of the participants, Asena and Eman, had an immigrant background: Asena's parents were from Turkey, while Eman's parents were from Bosnia-Herzegovina. Both of them were bilingual and spoke German and either Turkish or Bosnian, respectively. All students were fluent in academic German. Participants' ages ranged from 8 years 1 month to 10 years 1 month (M = 9 years 2 months). Their intelligence quotient (IQ) scores, as measured by the German Number Combination Test (ZVT; Oswald & Roth, 1987), ranged from 87 to 99 (M = 93.6). The students' fluency level according to the German Salzburg Reading and Orthography Test II (SLRT II; Moll & Landerl, 2010) was above average, with all students' scores exceeding the 75th percentile (range = 76-96%), indicating that the students were proficient decoders and fluent readers. However, the students' scores on the German Reading Comprehension Test for First to Six Graders (ELFE 1-6; Lenhard & Schneider, 2006), which measures students' ability to understand what they read, were remarkably low. Thus, all participants scored in the lowest third of their population with a percentage between 6 and 33%. According to the manual, the reliability of the ZVT varies between .84 and .97 (test-retest correlation). Comparisons between results from the ZVT and the Culture Fair Intelligence Test (CFT 3; Cattell, 1966) show r = .83. The retest-reliability of the SLRT II ranges between .80 and .97. Comparisons between results from the SLRT II and the Salzburg Reading Screening Instrument (SLS; Mayringer & Wimmer, 2003) amount to .75. For the ELFE 1-6, the test-retest correlation averages .91. Comparisons between results from the ELFE 1-6 and teacher appraisals amount to r = .70. Table 1 presents a summary of students' demographic information.

Table 1

Student Demographic Information (With Percentiles for Fluency and Comprehension Scores)

Student	Gender	Age (year-month)	IQ[a]	Fluency Score[b]	Comprehension Score[c]
Asena	F	9-8	90	96%	16%
Eman	M	8-6	96	83%	33%
Julia	F	8-1	87	76%	6%
Leon	M	10-1	99	88%	18%
Marvin	M	9-10	96	76%	11%

[a]ZVT: Number Combination Test (Oswald & Roth, 1987). [b]SLRT II: Salzburg Reading and Orthography Test (Moll & Landerl, 2010). [c]*ELFE 1-6: Reading Comprehension Test for First to Sixth Graders* (Lenhard & Schneider, 2006).

Interventionists. Two graduate students in special education from a university in the western part of Germany served as the interventionists and administered all instructional sessions across conditions. Prior to the start of the study, the first author trained the interventionists on the instructional procedures for teaching the use of the story mapping procedure. To ensure treatment fidelity, the interventionists were provided with a detailed script to follow and to assess their implementation of the strategy. In addition, the first author was in regular contact with them via e-mail and phone. Finally, during this intervention period, he held four formal research meetings to discuss issues related to implementation of the intervention in the schools.

Setting

The students were enrolled in two schools in North Rhine-Westphalia, Germany. Three of the students, Asena, Eman, and Julia, attended an inclusive elementary school in an outlying district of a major city, while Leon and Marvin were enrolled in a rural special school for slow learners. In both schools, the study took place in a room outside the students' classrooms during a daily period of independent class work.

Materials

Eighteen stories from three German storybooks (Wölfel, 1974, 2010a, 2010b) were used. The stories were short and modified to consist of exactly 150 words and contain all of the key story grammar elements. Ten story grammar comprehension questions were generated for each story, expressed in such a way that only one answer was possible. The level of reading difficulty of the pool of questions was assessed with 10 low-achieving students between 9 and 10 years of age. Based on this assessment, questions that were not answered correctly by at least five students were replaced or rephrased.

General Procedures

The study was conducted for 15 consecutive school days and 3 additional days that were evenly dispersed over 3 weeks following the intervention. Procedures identical to those used in the Grünke et al. (2013) study were implemented. All instructional sessions across conditions were carried out in a 1:1 format. During each lesson, the students read a story and provided written answers to 10 comprehension questions. Stories were presented in random order. The students did not receive any assistance or performance feedback on their answers to the comprehension questions.

Experimental Procedures

Baseline. During the baseline phase, the students silently read a story, rehearsed its content, and answered 10 comprehension questions within a 15-minute session. They were provided with a pencil, scratch paper, and a copy of the story. A timer was set to monitor the duration of the session. During silent reading of the story, the participants were allowed to consult any aid of their choice to memorize and make themselves familiar with the content (e.g., take notes, rehearse the information verbally, draw pictures). After the students finished reading and stated they were familiar with the story, the interventionists collected the copy of the story and any student-generated aids (e.g., notes, pictures). Next, the participants completed 10 comprehension questions related to the story grammar elements within the story. At the end of the 15-minute period, the students were asked to submit their responses to these questions to the interventionists.

Intervention. During the intervention phase, the participants were taught to use a story map using a procedure similar to Idol (1987), which consisted of three phases: a Model phase, a Lead phase, and a Test phase. Students read a story, completed a story map, and answered 10 comprehension questions in a 30-minute session. At the beginning of the intervention sessions, the students were provided with a pencil and a copy of a story. During the Model and Lead phase instructional sessions, the students also received a blank copy of a German version of a story map (see Figure 1 for an example).

During the Model phase, students were shown how to use a story map while reading a text. First, the interventionists sat next to the students, displayed a German version of the story map on a sheet of paper, and provided the students with a copy of a story and a blank story map. Next, the interventionists read the story aloud as the students followed along. While reading the story, the interventionists paused when a relevant story grammar element was identified in the text, filled out the appropriate parts of the story map, and asked the students to do the same on their own copy of the story map. Upon completion of the story, the participants turned in the reading passage and the completed story map to the interventionists.

In the Lead phase, the students read a story independently and completed a story map with minimal support from the interventionists. However, assistance was provided to scaffold the process, provide feedback on how the students were to complete the story map, answer story-related questions, remind students to be mindful of the key story grammar elements within the story, and to review their completed story map. If needed, the interventionists also assisted the students in finishing the story map. After completing the reading, the students turned in their copy of the story, filled out a story map, and answered 10 comprehension questions.

Finally, during the Test phase, the students read a story independently and completed their own story map on a piece of scratch paper, while the interventionists loosely monitored their work, answered questions pertaining to the story grammar elements, and provided support when students explicitly asked for help or if it was evident that they needed assistance to identify story grammar elements in the story.

All students completed two Model phase sessions. After the second lesson, they had reached a basic level of proficiency in the story mapping strategy and were able to move on to the next step, the Lead phase. For the Lead phase, Eman and Marvin received two instructional sessions, Asena received four, and Julia and Leon each received five. The criterion to advance from the Lead phase to the Test phase required the students to correctly complete the story maps with no assistance for two consecutive Lead phase sessions with 90% accuracy. All participants continued in the Test phase until they had completed the predetermined number of intervention sessions.

Maintenance. During the maintenance phase, procedures identical to those described in the baseline phase were implemented.

Experimental Design

A multiple-baseline design across participants (Baer, Wolf, & Risley, 1968) was used to determine the effects of the story mapping strategy. This approach demonstrates experimental control by systematically introducing the intervention in a time-lagged manner (Gast & Ledford, 2010). As mentioned above, introducing a random procedure into the design is an essential prerequisite for applying a randomization test, but it also strengthens the explanatory power of the whole study. Carrying on with baseline measurements until the variability in the data stabilizes may bias the design toward a particular intervention effect, thus compromising internal validity. The baseline data path may have shown more variability if baseline observations were allowed to continue. Moreover, several high and low random data points in the baseline might be mistaken for baseline stabilization (Todman, 2002). Introducing the intervention at random in a multiple-baseline design with a predetermined number of probes and a minimum number of baseline and intervention sessions controls for potentially systematic error and strengthens the internal validity of the findings (Dugard, 2013; Marascuilo & Busk, 1988).

The following specifications were established for the study:

1. A total number of 15 daily sessions were chosen for the baseline and the intervention sessions.
2. The baseline phase had to consist of at least three probes and the intervention phase had to consist of at least five probes, yielding eight possible intervention starting points, from the 4th to the 11th session (after the 3rd, 4th, 5th, 6th, 7th, 8th, 9th, or 10th baseline probe).
3. The starting point for each participant was randomly selected from the eight possible intervention starting points. Asena's intervention started after the fourth baseline probe, for Eman after the eighth, for Julia and Leon after the fourth, and after the fifth for Marvin.
4. To assess the continuation of the intervention effects, all the students received three maintenance probes after completion of the intervention phase.

Inter-Rater Reliability

Student responses to the comprehension questions were independently scored by the two interventionists for all probe sessions. A point was awarded for each question that was answered correctly. The maximum number of possible points earned in each probe was 10. Initial mean inter-rater agreement was 97%. Disagreements in scoring were extremely rare; however, when they happened, they were discussed and resolved by consensus to reach a 100% inter-rater reliability agreement.

Data Analysis

The data were analyzed through visual inspection examining the level, trend, and variability within and between phases (Gast & Spriggs, 2010). In addition, a one-tailed Monte Carlo randomization test for multiple-baseline designs across participants (AB) at a 0.05 significance level was applied to participants' baseline and intervention score data using a Microsoft® Excel macro downloaded from http://www.routledge.com/books/details/9780415886932/ (Dugard et al., 2012). As mentioned, our design involved an A, a B, and another A phase. However, there is no way to statistically analyze data from such a procedure using a randomization test (P. Dugard, personal communication, February 12, 2012). As a result, we had to limit ourselves to just considering the first A and the B phase for this part of the analysis.

The macro was set to generate 2,000 arrangements of the data at random. The sum of differences between intervention and baseline means was selected as the test statistic because the comprehension scores were expected to increase with the introduction of the intervention. With eight possible intervention starting points and five participants, there were $8^5 = 32{,}768$ arrangements of the data. Thus, the lowest possible *p*-value would be $1/32{,}768 = 0.00003$ if an exact

randomization test was conducted; with a Monte Carlo randomization test with 2,000 random samples, the lowest estimate p-value would be 0.0005. This created very favorable conditions for detecting an intervention effect in case it actually existed.

Finally, the improvement rate difference (IRD) for single-case research designs was calculated for all students (Parker, Vannest, & Brown, 2009). This effect size measures "… the difference in successful performance between baseline and intervention phases" (Alresheed, Hott, & Bano, 2013, p. 10). The IRDs were computed using the IRD calculator available at http://www.singlecaseresearch.org/calculators/ird.

Results

Table 2 and Figure 2 present the number of correctly answered comprehension questions during the baseline, intervention, and maintenance phases.

Table 2

Overview of Study Results per Phase

Student		Baseline	Intervention			Maintenance
			Model Phase	Lead Phase	Test Phase	
Asena	*N* (Probes)	4	2	4	5	3
	Scores	3; 0; 1; 1	4; 3	10; 8; 8; 6	9; 9; 8; 6; 10	10; 9; 6
	M	1.25	3.50	8.00	8.40	8.33
	IRD	-/-	0.50[1]	0.83[2]	0.90[3]	1.00[4]
Eman	*N* (Probes)	8	2	2	3	3
	Scores	2; 3; 0; 1; 3; 3; 1; 2	10; 10	10; 10	10; 9; 10	10; 8; 10
	M	1.88	10.00	10.00	9.67	9.33
	IRD	-/-	1.00[1]	1.00[2]	1.00[3]	1.00[4]
Julia	*N* (Probes)	4	2	5	4	3
	Scores	2; 0; 5; 1	3; 6	8; 9; 7; 6; 10	8; 8; 5; 10	10; 10; 8
	M	2.00	4.50	8.00	7.75	9.33
	IRD	-/-	0.50[1]	0.86[2]	0.75[3]	1.00[4]
Leon	*N* (Probes)	4	2	5	4	3
	Scores	4; 0; 5; 2	5; 7	9; 8; 10; 8; 8	10; 9; 8; 10	9; 9; 8
	M	2.75	6.00	8.60	9.25	8.67
	IRD	-/-	0.50[1]	0.86[2]	0.91[3]	1.00[4]
Marvin	*N* (Probes)	5	2	2	6	3
	Scores	2; 1; 3; 0; 2	10; 8	10; 10	10; 10; 10; 10; 10; 10	9; 10; 10
	M	1.60	9.00	10.00	10.00	9.67
	IRD	-/-	1.00[a]	1.00[b]	1.00[c]	1.00[d]

[a]Baseline vs. Model phase. [b]Baseline vs. Model + Lead Phase. [c]Baseline vs. Model + Lead + Test Phase. [d]Baseline vs. Maintenance Phase.

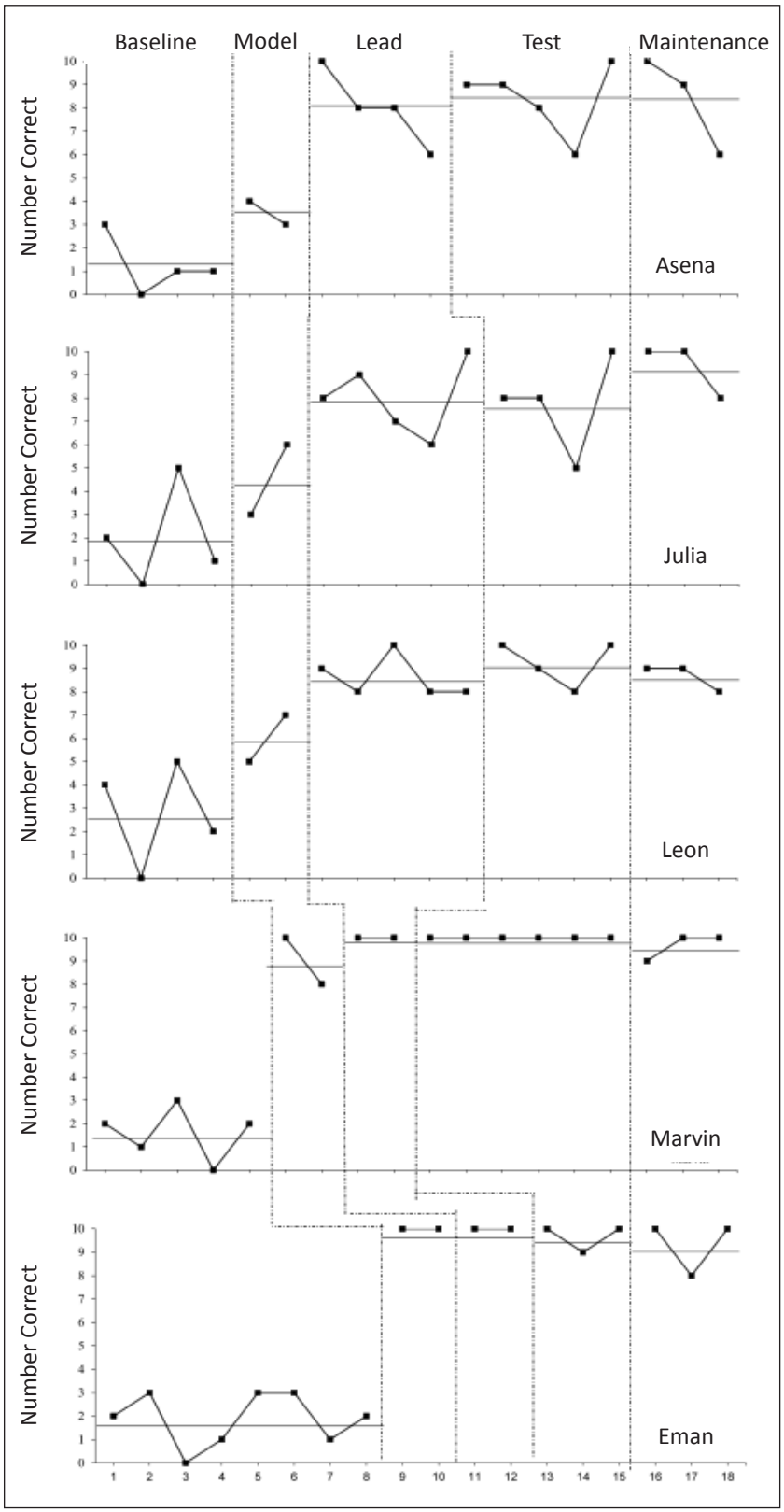

Figure 2. Number of correctly answered comprehension questions across baseline, intervention, and maintenance phases for Asena, Julia, Leon, Marvin, and Eman.

Results of the randomization test applied to the baseline and intervention phases showed that the differences between these two phases in the comprehension skills of the five students were statistically significant ($p < .001$; one-tailed). Furthermore, visual inspection of the data support the findings of the randomization test as all five students showed an increase in the number of story grammar elements identified in the story after the interventionists introduced the story mapping strategy. On average, during the baseline phase, participants were only able to answer fewer than three of the comprehension questions correctly, with mean scores ranging from 1.25 to 2.75. By contrast, in the intervention phase, students' overall mean comprehension scores ranged from 7.27 to 9.86. The average comprehension scores of Asena, Julia, Leon, and Marvin increased from the Model to the Lead phase as follows: from 3.50 to 8.00 for Asena, from 4.50 to 8.00 for Julia, from 6.00 to 8.60 for Leon, and from 9.00 to 10.00 for Marvin. The remaining student, Eman, scored 10 out of 10 during both the Model and Lead phases. From the Lead to the Test phase, the mean comprehension scores of Asena and Leon continued to improve, from 8.00 to 8.40 for Asena, and from 8.60 to 9.25 for Leon. However, a slight decrease in average performance was noted for Eman (from 10.00 to 9.67) and Julia (from 8.00 to 7.75). Alternatively, Marvin continued to answer all of the comprehension questions correctly in the Test phase. In the course of the maintenance phase, all students had average scores ranging from 8.33 to 9.67, with Eman, Julia, Leon, and Marvin scoring n 8-10 comprehension questions answered correctly.

Moreover, across all participants, IRD scores ranged from 0.50 to 1.00 between baseline and Model phase, and from 0.83 to 1.00 between baseline and Model phase + Lead phase. The overall IRD between baseline and intervention (Model phase + Lead phase + Test phase) ranged from 0.75 to 1.00. Between baseline and maintenance, IRD scores were 1.00 for all students (see Table 2). According to Alresheed et al. (2013), IRD scores between 0.70 and 0.75 are considered large or very large.

Overall, the results of the randomization test, the visual data inspection, and the IRD scores indicate that the use of a story map was an effective strategy for learning and acquiring the key story grammar elements in a story passage for five elementary students with LD. Finally, the IRD scores and visual analysis between baseline and maintenance phases for all students suggest that the effect of the story mapping strategy on the students' reading comprehension skills continued after the completion of the intervention.

Discussion

Over the last few years, the importance of SCR in identifying EBPs has increased, and researchers and practitioners alike now often resort to this kind of quality control

when evaluating the effects of a given intervention. However, in order to make sound instructional decisions, it is crucial to analyze and interpret a data set as objectively as possible. Conventional visual inspection often leaves too much to the discretion of the researcher in terms of drawing conclusions about the efficacy of an intervention. Thus, recent research has attempted to supplement traditional visual inspection of data from single-case research designs with statistical analysis (see Gage & Lewis, 2013, for a review).

The purpose of this study was to illustrate the use of a randomization test using a story map to promote the reading comprehension skills of five elementary students with LD. Results indicated that the participants increased the number of story grammar elements answered correctly from the baseline (overall $M = 1.90$) to the intervention (overall $M = 8.53$) and maintenance phases (overall $M = 9.07$). Moreover, the newly obtained skills were sustained after the instruction was withdrawn at levels near to those obtained in the Test phase during intervention. Based on the students' data, the results of the randomization test showed a significant difference between the baseline and the intervention mean scores at a 1% significance level, which rejects the null hypothesis of no intervention effect. Moreover, the two other procedures applied to measure the effectiveness of the intervention (i.e., visual inspection and effect size calculation) also suggested that story mapping is an effective strategy to recall and comprehend the key story grammar elements within a story passage. Together, the evidence suggests a large effect that positively impacted student comprehension outcomes related to story grammar.

Limitations

The study's findings show promising results; however, several limitations must be considered in this replication study. First, the sample size was small, which is a general limitation of SCR. The present study consisted of five elementary students with LD, which limits the generalizability of the findings. Second, differences between the duration of the baseline and maintenance sessions (15 minutes) and intervention sessions (30 minutes) may have positively affected students' reading comprehension scores during the intervention phase. Nevertheless, participants' comprehension levels in the maintenance phase were similar to those achieved during the Test phase. Third, inter-rater reliability measures were conducted by the two interventionists, which may have introduced a bias in scoring. It would have been more appropriate to have external observers independently score the students' answers to the comprehension questions. Fourth, no formal procedural reliability measures were conducted. However, the first author met on a regular basis with the interventionists to ensure that the procedures were consistently delivered as planned. Finally, although the randomization test was able to substantiate a strong intervention effect triggered by the treatment, an important limitation of this method of analyzing data from randomized SCR designs must be considered

in relation to slope effects. Wilbert (2014) was able to demonstrate that the randomization test is not sufficiently sensitive to slope effects. This limitation did not apply to our study, because the participants responded quickly to the intervention. However, in cases where the indicators of a treatment effect change more slowly, randomization tests might not be adequate for analyzing data from SCR designs. Thus, even though a significant effect of the story mapping strategy was observed in our study through the use of a randomization test, such an approach may not always be feasible for evaluating methods that elicit rather gradual responses.

Implications

In summation, the present study confirms insights from previous research on the positive effects of using a story map to improve the reading comprehension skills of students with LD. In addition, our findings also corroborate the results of Grünke et al. (2013) indicating that elementary students with moderate information-processing and poor comprehension skills, but with proficient decoding and fluency abilities can benefit from a rather short story mapping intervention of less than 12 sessions.

This entails some important implications for working with students comparable to the ones in our study. Specifically, it is possible to significantly assist students with average or above-average reading fluency but poor comprehension skills if key skills are targeted related to comprehension outcomes. Only about 10 lessons were needed to master the phase using the story mapping strategy. In most cases, such a short intervention can be embedded in typical reading instruction, thus preventing students from falling behind their classmates.

With regard to the randomization test, our study mainly focused on illustrating the technique. Thus, we demonstrated that this procedure can easily be implemented when conducting single-case analyses. In our case, the number of probes for the baseline and the intervention phase was only 15. Thus, the expenditure of time needed to conduct the experiment was minimal while still yielding a statistically significant result. As Dugard (2013) pointed out, the benefits of this method derive not only from the chance to verify statistically significant treatment effects, but the explanatory power of visual inspection also improves because the procedure requires the researcher to choose the intervention point at random.

Todman and Dugard (1999) noted that interpreting SCR data is problematic, leading to false conclusions when determining the differences between phases if no random assignment of treatments or conditions is used. The randomization test is one of the few methods that have the potency to validly detect statistically significant effects in data from SCR designs. Since this procedure can now easily be performed in familiar computational environments, it is hoped the approach will become more widely used among researchers who are trying to identify effective interventions.

References

Alresheed, F., Hott, B. L., & Bano, C. (2013). Single-subject research: A synthesis of analytic methods. *Journal of Special Education Apprenticeship, 2,* 1-18.

Al-Yagon, M., Cavendish, W., Cornoldi, C., et al. (2013). The proposed changes for DSM-5 for SLD and ADHD: International perspectives. *Journal of Learning Disabilities, 46,* 58-72.

Anderson, R. C., & Pearson, P. D. (1984). A schema-theoretic view of basic processes in reading comprehension. In P. D. Pearson (Ed.), *Handbook of reading research* (pp. 255-291). New York, NY: Longman.

Ausubel, D. P. (1960). The use of advance organizers in learning and retention of meaningful material. *Journal of Educational Psychology, 51,* 267-272.

Ausubel, D. P. (1968). *Educational psychology: A cognitive view.* New York, NY: Holt, Rinehart, & Winston.

Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 1,* 91-97.

Boulineau, T., Fore, C., Hagan-Burke, S., & Burke, M. D. (2004). Use of story mapping to increase the story-grammar text comprehension of elementary students with learning disabilities. *Learning Disability Quarterly, 27,* 105-121.

Brossart, D. F., Parker, R. I., Olson, E. A., & Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification, 30,* 531-563.

Bulté, I., & Onghena, P. (2013). The single-case data analysis package: Analyzing single-case experiments with R software. *Journal of Modern Applied Statistical Methods, 12,* 450-478.

Callahan, C. D., & Barisa, M. T. (2005). Introduction to the special issue of rehabilitation psychology: Issues in outcome measurement. *Rehabilitation Psychology, 50,* 5.

Campbell, J. M., & Herzinger, C. V. (2010). Statistics and single subject research methodology. In D. L. Gast (Ed.), *Single-subject research design in behavioral sciences* (pp. 417-453). Hillsdale, NJ: Lawrence Erlbaum.

Cattell, R. B. (1966). *Culture fair intelligence test: Scale 3.* Champaign, IL: IPAT.

Codding, R. S., Burns, M. K., & Lukito, G. (2011). Meta-analysis of basic-fact fluency interventions: A component analysis. *Learning Disabilities Research & Practice, 26,* 36-47.

Cook, B. G., Buysse, V., Klingner, J., Landrum, T. J., McWilliam, R. A., Tankersley, M., & Test, D. W. (2014). CEC's standards for classifying the evidence base of practices in special education. *Remedial and Special Education.* doi:10.1177/0741932514557271

Cook, B. G., & Cook, S. C. (2013). Unraveling evidence-based practices in special education. *The Journal of Special Education, 47,* 71-82.

Cook, B. G., Tankersley, M., & Harjusola-Webb, S. (2008). Evidence-based special education and professional wisdom: Putting it all together. *Intervention in School and Clinic, 44,* 105.

Cook, B. G., Tankersley, M., & Landrum, T. J. (2009). Determining evidence-based practices in special education. *Exceptional Children, 75,* 365-383.

Cook, B. G., Tankersley, M., & Landrum, T. J. (2013). *Evidence-based practices.* Bingley, United Kingdom: Emerald Insight.

Cook, B. G., Tankersley, M., Cook, L., & Landrum, T. J. (2008). Evidence-based practices in special education: Some practical considerations. *Intervention in School and Clinic, 44,* 69-75.

Davis, Z. T., & McPherson, M. D. (1989). Story map instruction: A road map for reading comprehension. *Reading Teacher, 43,* 232-240.

Dugard, P. (2013). Randomization tests: Are they what you need? *Insights on Learning Disabilities, 10,* 87-93.

Dugard, P., File, P., & Todman, J. (2012). *Single-case and small-n experimental designs: A practical guide to randomization tests.* New York, NY: Routledge.

Edgington, E. S. (1992). Nonparametric tests for single-case experiments. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis* (pp. 133-157). Hillsdale, NJ: Erlbaum.

Ferron, J. (2002). Reconsidering the use of the general linear model with single-case data. *Behavior Research Methods, Instruments, & Computers, 34,* 324-331.

Ferron, J., Foster-Johnson, L., & Kromrey, J. (2003). The functioning of single-case randomization tests with and without random assignment. *The Journal of Experimental Education, 71,* 267-288.

Gage, N. A., & Lewis, T. J. (2013). Analysis of effect for single-case design research. *Journal of Applied Sport Psychology, 25,* 46-60.

Gardill, M. C., & Jitendra, A. K. (1999). Advanced story map instruction: Effects on the reading comprehension of students with learning disabilities. *The Journal of Special Education, 33(*1), 2-17, 28.

Garthwaite, P. H., & Crawford, J. R. (2004). The distribution of the difference between two t-variates. *Biometrika, 91,* 987-994.

Gast, D. L., & Ledford, J. (2010). Multiple baseline and multiple probe designs. In D. L. Gast (Ed.), *Single-subject research methodology in behavioral sciences* (pp. 276-328). New York, NY: Routledge.

Gast, D. L., & Spriggs, A. D. (2010). Visual analysis of graphic data. In D. L. Gast (Ed.), *Single-subject research methodology in behavioral sciences* (pp. 199-233). New York, NY: Routledge.

Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children, 71,* 149-164.

Grimm, J., & Grimm, W. (2013). *The frog prince.* New York, NY: NorthSouth.

Grünke, M., & Calder Stegemann, K. (2014). Using count-bys to promote multiplication fact acquisition for a student with mild cognitive delays: A case report. *Insights on Learning Disabilities, 11,* 117-128.

Grünke, M., Wilbert, J., & Calder Stegemann, K. (2013). Analyzing the effects of story mapping on the reading comprehension with low intellectual abilities. *Learning Disabilities: A Contemporary Journal, 11,* 51-64.

Harbst, K. B., Ottenbacher, K. J., & Harris, S. R. (1991). Interrater reliability of therapists' judgments of graphed data. *Physical Therapy, 71,* 107-115.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71,* 165-179.

Idol, L. (1987). Group story mapping: Comprehension strategy for both skilled and unskilled readers. *Journal of Learning Disabilities, 20,* 196-205.

Idol, L., & Croll, V. J. (1987). Story-mapping training as a means of improving reading comprehension. *Learning Disability Quarterly, 10,* 214-229.

Janosky, J. E., Al-Shboul, Q. M., & Pellitieri, T. R. (1995). Validation of the use of a nonparametric smoother for the examination of data from a single-subject design. *Behavior Modification, 19,* 307-324.

Jitendra, A. K., & Gajria, M. (2011). Reading comprehension instruction for students with learning disabilities. *Focus on Exceptional Children, 43,* 1-16.

Johnson, L., Graham, S., & Harris, K. R. (1997). The effects of goal setting and self-instruction on learning a reading comprehension strategy: A study of students with learning disabilities. *Journal of Learning Disabilities, 30,* 80-92.

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods, 15,* 124-144.

Lee, O.-I., & Kim, J. (2013). Reading strategies for students with learning disabilities: Meta-analysis of single subject researches. *Journal of Convergence Information Technology, 8,* 429-439.

Lenhard, W., & Schneider, W. (2006). *Reading comprehension test for first to six graders (ELFE 1-6).* Göttingen, Germany: Hogrefe.

Manolov, R., Arnau, J., Solanas, A., & Bono, R. (2010). Regression-based techniques for statistical decision making in single-case designs. *Psicothema, 22,* 1026-1032.

Marascuilo, L. A., & Busk, P. L. (1988). Combining statistics for multiple-baseline AB and replicated ABAB designs across subjects. *Behavioral Assessment, 10,* 1-28.

Mastropieri, M. A., Scruggs, T. E., Mills, S., et al. (2009). Persuading students with emotional disabilities to write fluently. *Behavioral Disorders, 35,* 19-40.

Matson, J., Turygina, N., Beighleya, J., & Matson, M. (2012). Status of single-case research designs for evidence-based practice. *Research in Autism Spectrum Disorders, 6,* 931-938.

Mayringer, H., & Wimmer, H. (2003). *Salzburg reading screening instrument (SLS).* Göttingen, Germany: Hogrefe.

Moll, K., & Landerl, K. (2010). *Salzburg reading and orthography test (SLRT II).* Göttingen, Germany: Hogrefe.

Odom, S. L., Brantlinger, E., Gersten, R., Homer, R. H., Thompson, B., & Harris, K. R. (2005). Research in special education: Scientific methods and evidence-based practices. *Exceptional Children, 71,* 137-148.

O'Donnell, A. M., Dansereau, D. F., & Hall, R. H. (2002). Knowledge maps as scaffolds for cognitive processing. *Education Psychology Review, 14,* 71-86.

Oswald, W. D., & Roth, W. (1987). *Number combination test (ZVT).* Göttingen, Germany: Hogrefe.

Parker, R. I., Vannest, K. J., & Brown, L. (2009). The improvement rate difference for single-case research. *Exceptional Children, 75,* 135-150.

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy, 42,* 284-299.

Regan, K. S., Mastropieri, M. A., & Scruggs, T. E. (2005). Promoting expressive writing among students with emotional and behavior disturbance via dialogue journals. *Behavioral Disorders, 31,* 33-50.

Riley-Tillman, T. C., & Burns, M. K. (2009). *Evaluating educational interventions: Single-case design for measuring response to intervention.* New York, NY: Guilford.

Scruggs, T. E., Mastropieri, M. A., & Regan, K. S. (2006). Statistical analysis for single subject research designs. *Advances in Learning and Behavioral Disabilities, 19,* 33-53.

Stagliano, C., & Boon, R. T. (2009). The effects of a story mapping procedure to improve the comprehension skills of expository text passages for elementary students with learning disabilities. *Learning Disabilities: A Contemporary Journal 7,* 35-58.

Taylor, L. K., Alber, S. R., & Walker, D. W. (2002). A comparative analysis of a modified self-questioning strategy and story mapping on the reading comprehension of elementary students with learning disabilities. *Journal of Behavioral Education, 11,* 69-87.

Todman, J. (2002). Randomisation in single-case experimental designs. *Advances in Clinical Neuroscience and Rehabilitation, 2,* 18-19.

Todman, J., & Dugard, P. (1999). Accessible randomization tests for single-case and small-n experimental designs in AAC research. *Augmentative and Alternative Communication, 15,* 69-82.

Torres, C., Farley, C. A., & Cook, B. G. (2014). A special educator's guide to successfully implementing evidence-based practices. *Teaching Exceptional Children, 47,* 85-93.

Wade, E., Boon, R. T., & Spencer, V. G. (2010). Use of Kidspiration software to enhance the reading comprehension of story grammar components for elementary-age students with specific learning disabilities. *Learning Disabilities: A Contemporary Journal, 8,* 31-41.

Wilbert, J. (2014, August). *Which technique is appropriate for analyzing single-case AB designs?* Paper presented at the biannual Special Educational Needs Conference of the European Association for Research on Learning and Instruction (EARLI), Zürich, Switzerland.

Wölfel, U. (1974). *Neunundzwanzig verrückte Geschichten* [Twenty-nine crazy stories]. Stuttgart, Germany: Hoch.

Wölfel, U. (2010a). *Achtundzwanzig Lachgeschichten* [Twenty-eight laugh stories]. Stuttgart, Germany: Thienemann.

Wölfel, U. (2010b). *Siebenundzwanzig suppengeschichten* [Twenty-seven soup stories]. Stuttgart, Germany: Thienemann.

Zheng, X., Flynn, L. J., & Swanson, H. L. (2013). Experimental intervention studies on word problem solving and math disabilities: A selective analysis of the literature. *Learning Disability Quarterly, 36,* 97-111.