Application of a Single-Case Intervention Procedure

To Assess the Replicability of a Two-Component Instructional Strategy

Yooyeun Hwang, Hope College

Joel R. Levin, University of Arizona

Abstract

A dual-component single-case multiple-baseline design and statistical analysis was implemented to assess the replicability of instructional-strategy effects that have been well established in previous conventional randomized "group" intervention research. The 15-week intervention study examined the efficacy of a sequentially presented pictorial "mnemonic numeric" strategy designed to help seven 11- and 12-year-old children remember the dates (centuries and decades) of various 18th, 19th, and 20th century inventions. Two different single-case multiple-baseline randomization-test procedures were applied to confirm a predicted set of differentiated experimental outcomes. Suggestions were provided for modifying and improving the methods' suitability for single-case educational intervention researchers.

Application of a Single-Case Multiple-Baseline Intervention Procedure

To Assess the Replicability of a Two-Component Instructional Strategy

It is widely accepted that randomization (i.e., random assignment) and experimental control represent the defining characteristics of an internally valid, scientifically credible, educational intervention research study (e.g., Levin, 1994; Shadish, Cook, & Campbell, 2002). Similarly, random sampling and replication/reproducibility are the hallmarks of externally valid educational intervention research, in that without those two critical ingredients across-study generalizations are not warranted (Shadish et al., 2002). Concerning replication, a sad state of affairs in the social and behavioral sciences is that too many seemingly "promising" new findings do not stand up to the replication test (e.g., Ioannidis, 2005), and which, in the field of psychology has given rise to what is known as the Reproducibility Project (Open Science Collaboration, 2012). In those instances where replication attempts of a conventional randomized "large-sample" intervention study (hereafter referred to as conventional "group" research) are made, they are typically made in the context of a similar randomized "group" study.

In the present investigation, our primary objective is to determine whether an instructional intervention that has proven to be effective in conventional randomized "group" research studies will produce similar benefits through implementation of randomized, carefully conducted (but on a much smaller-scale) "single-case" intervention methods (see, for example, Kratochwill & Levin, 2014). That is, we will examine the research-replicability potential of a scientifically sound methodological approach that requires a very small number of participants. Our focal strategy is a dual-component mnemonic (memory-enhancing) technique for helping children remember the dates of selected inventions from the 18th, 19th, and 20th centuries (e.g., Hwang & Levin, 2002; Hwang, Renandya, Levin, Levin, Glasman, & Carney, 1999).

In recent years, the implementation of single-case intervention designs as a *bona fide* form of academic inquiry has been gaining traction in the fields of education and psychology, among others (see, for example, Evans, Gast, Perdices, & Manolov, 2014; Kratochwill, Hitchcock, Horner, Levin, Odom, Rindskopf, & Shadish, 2013; Kratochwill & Levin, 2014). What has elevated the scientific credibility of such designs is their incorporation of various forms of randomization that characterize conventional experimental and randomized controlled trials research, incorporations serving to enhance the designs' internal validity (Kratochwill & Levin, 2010, 2014). Within the single-case intervention design repertoire is the multiple-baseline design, which is among the most popular, versatile, and scientifically credible designs with respect to its ability to document a direct link between interventions and outcomes (e.g., Horner & Odom, 2014; Levin, 1992; Levin, Ferron, & Gafurov, 2018). Along with the recent methodological enhancements of single-case intervention designs, in general, and multiple-baseline designs, in particular, are concomitant enhancements in the statistical-conclusion validity of the data-analysis tools that are implemented to assess single-case research outcomes (e.g., Ferron & Levin, 2014; Ferron, Moeyaert, Van den Noortgate, & Beretvas, 2014; Shadish, 2014).

Let us first consider a traditional multiple-baseline design that satisfies the Institute of Education Science's (IES's) What Works Clearinghouse (WWC) Standards that qualify it as an "acceptable" single-case design (Kratochwill et al., 2013). The design must include at least three cases, or "tiers" (Barton & Reichow, 2012), with the tiers represented by different cases (either individuals or aggregates), settings/contexts, or measures (behavioral, cognitive, etc.). Within this "interrupted time-series design" framework, there is a baseline (A) phase consisting of multiple sequential observations, followed by an intervention (B) phase also consisting of multiple sequential observations, with the intervention introduced in a systematically staggered fashion across the tiers.

The design is outlined in Table 1, where the As and Bs within the table designate the individual baseline and intervention observations, respectively, that are collected over the study's 14 time periods. This hypothetical design includes four cases and it is assumed that the cases have been randomly assigned to the design's four tiers. Each case receives a minimum of 4 baseline observations, a minimum of 4 intervention observations, and a two-observation staggered introduction of the intervention between successive cases. Documenting evidence of an "intervention effect" (i.e., that there is a direct link between the intervention and a change in the targeted outcome measure) requires that: (a) the hypothesized change in the outcome measure is temporally coincident with the introduction of the intervention for each successive case; and (b) the change is evident *only* in the case for which the intervention has been introduced and not in any of the remaining cases that are still in the baseline phase. So, for example, if had been hypothesized that there would be an overall baseline-to-intervention-phase increase in the mean outcome measure (i.e., an increase in "level"), then each case should exhibit an increase during the point at which—and *only* during the point at which—the intervention has been introduced to that case. When conceptualized in this manner, the multiple-baseline design receives high internal-validity marks with respect to ruling out potential extraneous factors (i.e., factors other than the intervention *per se*) that could plausibly account for an obtained across-cases A-to-B-phase increase in level (Levin, 1992).

This brings us to the present study's two-component, two outcome measures, multiple-baseline design and analysis. In the single-case literature, a two-component (B and C) intervention design has sometimes been embedded within an AB design format. For example, instead of an alternating six-phase baseline-intervention ABABAB design, one could have an alternating baseline (A), two-intervention (B and C) design, such as ABACA(B+C) – see, for example, Kratochwill & Levin (2010). In the present multiple-baseline design, following a staggered baseline phase, seven 11- and 12-year-old children were presented a two-stage, sequentially

introduced, pictorial "mnemonic numeric" strategy to help them remember the dates (centuries and decades) of various inventions from the 18th, 19th, and 20th centuries (see, for example, Hwang & Levin, 2002). In Stage 1, the children were taught a strategy for remembering the inventions' centuries (Intervention B) but not their decades; and then in Stage 2, a decade-remembering strategy (Intervention C) was added to the century-remembering strategy. Throughout the 15-session multiple-baseline study, in each session the children were tested on their memory for both the centuries and the decades of the studied inventions.

Hypotheses Tested in the Present Study

Consistent with the logic of the multiple-baseline design, along with Campbell and Fiske's (1959) discriminant validity notions, it was hypothesized that with the Stage 1 staggered introduction of the mnemonic century strategy (B) following the A (baseline) no-strategy phase, there would be substantial improvement in the children's performance on the century-memory outcome measure (Outcome X) but little or no improvement on the decade-memory measure (Outcome Y). This gives rise to the following two Stage 1 hypotheses:

Hypothesis 1a: With the staggered introduction of the mnemonic century strategy, there *will be* an A phase to B phase improvement in the children's performance on the century outcome measure: specifically, B > A.

Hypothesis 1b: With the staggered introduction of the mnemonic century strategy, there *will not be* an A phase to B phase improvement in the children's performance on the decade measure: specifically, B ≈ A.[1]

Conversely, it was hypothesized that with the Stage 2 introduction of the mnemonic decade strategy (C) following the mnemonic century strategy phase (B), there would be substantial improvement in the children's performance on the decade-memory outcome measure (Outcome Y) but little or no improvement on the century-memory measure (Outcome X). Accordingly, we have the two following Stage 2 hypotheses:

Hypothesis 2a: With the introduction of the mnemonic decade strategy, there *will not be* an A/B to B+C phase improvement in the children's performance on the century outcome measure: specifically, B+C ≈ A/B.

Hypothesis 2b: With the introduction of the mnemonic decade strategy, there *will be* an A/B to B+C phase improvement in the children's performance on the decade outcome measure: specifically, B+C > A/B.

In addition, a 4-item new vocabulary-learning task was administered in each session to afford a more distal discriminant-validity measure assessment of the focal mnemonic inventions strategy. Two hypotheses were formulated for children's performance on the vocabulary task. Specifically:

Hypothesis 1c: In Stage 1, with the staggered introduction of the mnemonic century strategy, there *will not be* an A phase to B phase improvement in the children's performance on the vocabulary outcome measure: specifically, B ≈ A.

Hypothesis 2c: In Stage 2, with the introduction of the mnemonic decade strategy, there *will not be* an A/B phase to B+C phase improvement in the children's performance on the vocabulary outcome measure: specifically, B+C ≈ A/B.

One source of support for these hypotheses would consist of visual analyses that satisfy the WWC "evidence criteria" for a multiple-baseline design (Kratochwill et al., 2013), whereas another source would be derived from randomization statistical tests applied to the two-component multiple-baseline data (e.g., Levin et al., 2018).

We now present a description of the intervention study that was conducted here. Journal space limitations do not permit a complete account of the methods but details about the study's participants, materials, and procedures can be obtained from either author on request.

Method

Participants

The participants in this multiple-baseline study were a convenience sample of three fifth-grade and four sixth-grade elementary school students, consisting of two girls and five boys. Following the study's approval by the College's IRB, we recruited students from both a summer-school program and during the regular school year with the consent of teachers, parents, and children. Each of the children participated in several 15- to 20-minute individual sessions spread over a 4- to 6-week period. Six of the seven children completed all the study's 15 sessions but owing to a late start, one child completed only 13 sessions. The present multiple-baseline design is "nonconcurrent" because the children participated at different points in time during both the summer and the regular school year.

Design and Materials

Each of the seven participants received three sequentially administered learning strategies, in a staggered across-cases fashion: baseline control (A), century mnemonic (B), and century-plus-decade mnemonic (B+C). The students were randomly assigned to the seven tiers of the multiple-baseline design, with a one-session stagger between tiers. Given a fixed total number of 15 sessions for each child, it was decided that each child should receive a minimum of three sessions representing each design component (A, B, and B+C). This resulted in the child in the design's first tier receiving three A sessions and the child in the seventh tier receiving nine A sessions. All seven students received three B sessions. Finally, the first child received nine B+C sessions and the seventh child received three B+C sessions. Thus, across the seven children, the number of A (Baseline Control) sessions and outcome measures ranged from 3 to 9, the number of B (Century Mnemonic) sessions and outcome measures were 3 for all children, and the number of B+C (Century + Decade Mnemonic) sessions and outcome measures ranged from 3 to 9. In every session, the children first studied a list of four new inventions items (and four new vocabulary items, described below) and were then tested on them.

All materials were printed on 8-1/2" by 11" sheets of paper. The study materials contained line drawings and text, and the test materials contained only text. The content of both the study and test materials was read aloud to the students by a single experimenter.

Invention materials. Sixty inventions were selected from internet searches, with 20 apiece representing three different centuries (specifically, the 18th, 19th and 20th centuries). In addition, within each century, 20 inventions were chosen from the 10 different decades (i.e., the 00s, 10s, 20s . . . 90s), which included between one and three inventions from the same decade. In each session, the individually treated students were taught four new inventions and their dates (i.e., century and decade), with the specific inventions presented in the same order (within and across sessions) for all seven children. The study materials were prepared in three different formats, as are now briefly described.

1. Baseline control condition. During the A phase, the inventions were illustrated with line drawings and were labeled with the invention names and dates (consisting of the century and the decade) and without any additional objects or background depicted. For example, the wrench was invented in the 1830s.

2. Century mnemonic condition. During the B phase, each invention line drawing was accompanied by an illustration of people from their time periods, which was called the century setting. Specifically, royalty was used to represent the 18th century, cowboys to represent the 19th, and astronauts to represent the 20th.

To connect each invention with its date, interactive pictures were created by combining the invention and the century-setting representations. For the previous 1830s wrench example, and with cowboys representing the 19th century, the illustration depicted a cowboy fixing a wheel on his covered wagon with a wrench. In addition, the following verbal description was included as a short caption underneath the picture: "A cowboy is fixing his wagon using a *wrench*."

3. Century-plus-decade mnemonic condition. During the B+C phase, the students were taught a more complex mnemonic method for remembering both the century and the decade, which built on the century mnemonic strategy that the students had already learned during the B phase (see also Hwang & Levin, 2002). Specifically, in addition to the century setting, 10 months of the year were used for learning the 10 different decades, starting with December for number 0 and going from January for number 1 through September for number 9. With the exception of December, which is 0, each month matches the number with which it is usually associated (1 = January, 2 = February, 3 = March, and so on). Each month was then used to represent each decade. The 00s decade was December, the 10s decade was January, the 20s decade was February, the 30s decade was March, etc. To make each month picturable, we chose a familiar seasonal setting to represent it: the 00s decade, December, was represented by a Christmas scene (Santa Claus, presents, snow); the 10s decade, January, was represented by a New Year's Eve scene (party hats, decorations, a clock); the 20s decade, February, was represented by a Valentine's Day scene (hearts, lovers, sweets); the 30s decade, March, was represented by a Windy Day scene (the wind, rain, umbrellas); and so on.

To connect the inventions and their dates, interactive illustrations were created by combining the pictures of an invention, its century setting (i.e., people), and its seasonal setting (i.e., decade). For the previous 1830s wrench example, the B+C mnemonic picture included wrenches, a cowboy for the 1800s, and a windy March day scene representing the 30s decade. Specifically, the mnemonic picture showed the wind blowing away a cowboy's wrenches while he is trying to fix a wheel on his wagon. In addition, each picture was described by a short caption appearing under it, as for this example: "A cowboy is fixing his wagon and the wind is blowing away his *wrenches*."

Vocabulary study materials. Sixty unfamiliar and infrequently used vocabulary words were selected for the study materials. Each word and its definition (e.g., *corsair*, meaning pirate) were printed above a picture

that illustrated the meaning of the vocabulary. For example, the picture for *corsair* was an image of a pirate. In each session, the students were given four new words and their meanings, which were presented in the same order (within and across sessions) for all seven students.

Filler task materials. The students engaged in a word-search "filler" task after studying each session's target information (four inventions and dates, four vocabulary words) and before taking the inventions and vocabulary memory tests. The word-search task was administered to increase the interval length between the session's study and test portions, thereby helping to mitigate any short-term memory effects on the memory tests.

Instructions

A set of standardized instructions was developed for studying inventions and vocabulary words in the present multiple-baseline components (A, B, and B+C). For each component, there was one set of instructions for the inventions task and another set of instructions for the vocabulary task. For the baseline control component (A), the students were then informed that they would be presented four different inventions and their dates and asked to remember the century and the decade using whatever method worked best for them. For the century mnemonic component (B), the instructions explained how the students should use the pictures of people from the three different time periods to remember the inventions' centuries but that they still needed to memorize the decades, using whatever method worked well for them. For the century-plus-decade mnemonic component (B+C), the instructions first reminded the students how they had been using the mnemonic century settings and then explained how to add the monthly seasonal settings to represent the decade. For all components, the instructions informed the students that later they would be tested on the dates (centuries and decades) of those four inventions. In the second part of the session for each component (A, B, and B+C), the experimenter explained that the students would learn some unfamiliar English words and

their meanings, using whatever method worked best for them. The students were again told that after studying four new vocabulary words, they would be tested on their memory for those words' meanings.

Tests

Immediate tests were created to assess students' memory for the dates of the inventions and the meanings of the unfamiliar words that had been presented during the study portion of the session. The delayed inventions test assessed the students' memory for the inventions and their dates that were presented during the previous session. The primary reason that a delayed test was implemented here was that the true magnitude of the mnemonic invention strategy's effects likely would be attenuated on the immediate test, resulting from students' ceiling-level performance on the session's 4-item century-memory measure, as was recently reported by Hwang et al. (2016).

Procedure

The procedural format for Session 1 was structured in the following three segments. First, the students were presented four inventions and their dates, followed by four new vocabulary words. The experimenter attempted to present each item to students for the same amount of time both within and across the study's three phases (A, B, and B+C). Students then received the word-search filler task. Finally, the immediate memory test for vocabulary items was administered, followed by the immediate test for inventions. The longer interval between study and test of the inventions items than of the vocabulary items was established purposely to eliminate as much short-term memory carryover as possible for the inventions measure, the primary outcome of interest in the present study. In contrast, the vocabulary task was included mainly to afford a discriminant validity measure assessment of the focal mnemonic inventions strategy. For each subsequent session (Sessions 2-15) the students were first given a four-item delayed test to assess their memory of the inventions' dates that they had learned during the previous session.

All children received between 12 and 14 delayed tests. Most delayed tests were one-day-delayed tests. However, in the first session of the week, students were assessed on their memory for the last session of the previous week, which was typically a four-day-delayed test. However, there were a few 5- to 7-day-delayed tests and one 11-day-delayed test, resulting from a school holiday, absences, and scheduling conflicts. Analyses of the data with and without the three 7-day-delayed scores and the one 11-day-delayed score yielded the same statistical conclusions and so those four scores are included in the analyses reported here.

After each delayed inventions test, the students were presented new invention items to study, followed by four new vocabulary words and their meanings, according to the previously described procedure. Then, the word-search filler task was administered, followed by an eight-item memory test: the four new vocabulary words items followed by the four new inventions that had been presented in the current session.

Results

Ordinarily a single-case researcher would conduct a traditional "visual analysis" of the data (see, for example, Horner & Spaulding, 2010), as was done with a single-case crossover design and randomization test in a study that that was a forerunner to this one (Hwang et al., 2016). A traditional visual analysis is not reported here, however, because our major purpose is to implement a statistical approach for assessing discriminant validity hypotheses in the context of a two-component multiple-baseline design, through the application of recently developed single-case randomization tests (Levin et al., 2018).

Although a variety of statistical procedures have been proposed to analyze the data from single-case intervention designs, including the multiple-baseline design—see, for example, Hedges, Pustejovsky, & Shadish (2013); Kratochwill & Levin (2014); and Shadish et al. (2014)—here, we focus on what are known as nonparametric "randomization tests" (e.g., Dugard, File, & Todman, 2012; Edgington & Onghena, 2007) because of a number of positive features associated with such tests (Ferron & Levin, 2014). Randomization

statistical tests differ from conventional statistical tests in that whereas the latter's hypotheses are directed at specific population *parameters* (for example, a difference in two population means), the former's hypotheses are directed at differences between population *distributions*. To yield conclusions about specific parameters in those populations (e.g., that there is a difference in the population means), a test statistic is formulated that is sensitive to the hypothesis of interest (e.g., the difference between the B-phase and A-phase means). Then, if the randomization test proves to be statistically significant, a logical inference is made that the difference in the two population distributions is attributable (at least in part) to the outcome associated with the defined test statistic.

The multiple-baseline randomization-test program in Gafurov and Levin's (2018) freely available *ExPRT* statistical package was used for all the present analyses. In that the invention-task century and decade measures produced identical patterns on the immediate and delayed inventions tests (as did their associated statistical conclusions), we will consider only the delayed-test results because the patterns are more visually discernible in the to-be-presented graphs. No delayed tests on the vocabulary measure were given and so only immediate vocabulary test results are presented.

Two sets of analyses were conducted on all measures in the context of a multiple-baseline design with the intervention introduced in a staggered fashion across children. The Stage 1 analyses examined the children's performance-level change from the baseline phase (A) to the first mnemonic invention strategy (i.e., the mnemonic century) phase (B). The Stage 2 analyses examined the children's level change from the combined baseline and century strategy phases (hereafter designated as A/B) to the second strategy (i.e., the mnemonic decade) phase (B+C). Our reason for combining the A and B phases in these analyses is discussed below.

Each of the six previously presented hypotheses (Hypotheses 1a, 1b, and 1c; 2a, 2b, and 2c) was statistically assessed based on a Type I error probability (α) of .05 with a one-tailed alternative (i.e., B > A in Stage 1 and B+C > A/B in Stage 2). In addition to providing statistical decisions and significance probabilities (*p*-values) for each tested hypothesis, *EXPRT* provides two effect-size measures, *d* and NAP. The *d* measure indicates how far apart the first and second compared phase means are in standard deviation units (based on the first phase's standard deviation)—see Busk & Serlin (1992). The NAP measure (non-overlap of all pairs), and the one reported here, reflects the degree of non-overlap between the outcome observations from the two compared phases, with the resulting rescaled or "adjusted" NAP ranging from 0 (complete overlap) to 1 (complete non-overlap)—see Parker, Vannest, & Davis (2014); as well as Gafurov and Levin's (2018) *ExPRT* "User Instructions." Other nonoverlap measures, such as PAND and Tau-U (as detailed by Parker et al., 2014) could have been computed instead of NAP, but the *ExPRT* package provides NAP because of its intuitively appealing interpretation as well as its direct connection to nonoverlap measures in other scientific disciplines.

Stage 1 Analyses of the Delayed Century and Decades Measures

For the Stage 1 analyses, each child's baseline (A) observations were compared with the observations commencing with the introduction of (and extending through the application of) the mnemonic century strategy (B) – see Table 2. Recall that no delayed tests were administered in Session 1 and that the mnemonic century strategy (B) was not introduced in a session until the delayed memory test from the previous session had been administered. Thus, Child 1's A (baseline phase) delayed tests were administered at the beginning of Sessions 2, 3, and 4 and the B (mnemonic century phase) delayed tests were administered at the beginning of Sessions 5, 6, and 7. For Child 2, the two sets of delayed tests were given in Sessions 2, 3, 4, and 5 (A) and in Sessions 6, 7, and 8 (B), respectively; and so on. The Wampold-Worsham (1986) randomization test procedure was the preferred statistical method to analyze the data, but because the lengths of the children's series

differed in unacceptable ways (as is described by Gafurov & Levin, 2018), the somewhat less powerful modified Revusky (1967) randomization-test procedure (Levin et al., 2018) had to be used for the Stage 1 analyses instead.[2]

The results, presented in Figure 1, are straightforward and easy to describe. For the inventions items in the Stage 1 analyses, previously presented Hypothesis 1a specifies that there would be century-memory improvement coincident with the introduction of the century strategy that the children were taught. In Figure 1 it may be seen that with the exception of Child 1, whose century memory scores were perfect both before (A phase) and after (B phase) the mnemonic century strategy was administered: (1) each child's B-phase scores are at a higher level than his/her A-phase scores; and (2) that improvement basically occurs at the point where the strategy was introduced — consistent with the staggered intervention logic of the multiple-baseline design. The modified Revusky test of the intervention effect was statistically significant, $p = .01$, and the associated average adjusted NAP measure (across the 7 children) was equal to 0.74. Accordingly, Hypothesis 1a was statistically supported.

At the same time, Hypothesis 1b specifies that on the Stage 1 decade-memory measure, there would be no improvement because the children had not yet been taught the mnemonic decade strategy. The results are presented in Figure 2, where quite a different pattern can be seen than that for the century-memory measure in Figure 1. Specifically, no consistent performance-level improvement is apparent across the seven children, with the children remembering nonstatistically more decades in the A (baseline) phase than in the B (century strategy) phase and an average adjusted NAP of .06, favoring A > B. Accordingly, Hypothesis 1b was supported.

Stage 2 Analyses of the Delayed Century and Decades Measures

Ideally, to conduct a multiple-baseline comparison of the mnemonic decade strategy (C) phase and mnemonic century strategy (B) phase observations, we would have built a stagger into the B phase. Unfortunately, we could not do so in the present study because that would have required many more sessions than we were allowed. So, in our analysis based on seven children and 15 sessions, we had to do the next best thing. Specifically, we conducted a Stage 2 comparison between the B+C phase and the combined A and B phases. As a result, an analysis of the century measure includes baseline observations in addition to century-strategy observations, which poses a challenge to discriminant validity Hypothesis 2a (namely, not detecting a Stage 2 strategy effect on the century measure). That is because although the "intervention" phase has added a C (mnemonic decade) component to the B (mnemonic century) component, the "baseline" phase contains both A (baseline) and B (mnemonic century strategy) observations. Consequently, for many of the A/B outcome measures the children had already been taught the mnemonic century strategy and so we do not a have a "pure" test of the comparison of interest, the C phase vs. the B phase. The implications of that "impurity" are considered in the Discussion section.

That said, with a converse argument to the one applied to the Stage 1 hypotheses, in the Stage 2 analysis for the century-memory measure Hypothesis 2a specifies that there will be little or no improvement in the children's memory performance; and for the decade-memory measure, Hypothesis 2b specifies that there will be performance improvement. Because the same series lengths (comprised of 14 delayed tests) were available for all children here, it was possible to conduct Wampold-Worsham randomization tests on these data. Unfortunately, however, the century-memory results presented in Figure 3 are compromised by obvious ceiling effects for six of the seven children and so their implications for Hypothesis 2a (to be considered in the Discussion section) are equivocal and hence the Hypothesis 2a "Support?" entry in Table 3 is

accompanied by a "?". Given that "equivocal" caveat, the randomization test results were not statistically significant, p = .88, and an average adjusted NAP of .39, consistent with the specifications of Hypothesis 2a.

In contrast, the decade-memory results in Figure 4 reveal clear A/B to B+C phase improvements in the children's performance that are generally coincident with the introduction of the Stage 2 mnemonic decade strategy. The randomization test applied to these data produced a p-value of .001, with an average adjusted NAP of .80, thereby providing support for Hypothesis 2b.

Analyses of the vocabulary measure. For the vocabulary items, Hypotheses 1c and 2c indicate that there will be no statistical improvement in the children's performance in either the Stage 1 or Stage 2 analyses. That is because: (1) the children were not taught a strategy for learning the vocabulary items and so that task provides additional, more distal, discriminant-validity support for the targeted efficacy of the present mnemonic century and decade strategies; and (2) the vocabulary items were included simply as "filler" items to increase the difficulty of the focal inventions task. The results of the Stage 1 (modified Revusky procedure) and Stage 2 (Wampold-Worsham procedure) analyses are not depicted here in a Figure but are summarized in Table 3. As can be seen in that table, in neither the A to B phase Stage 1 analysis nor the A/B to B+C phase Stage 2 was there any suggestion of an intervention effect on the vocabulary items. Specifically, in the Stage 1 analysis, the children's average levels of recall for the four vocabulary items were 3.40 vs. 3.42 during the A and B phases, respectively, and 3.42 vs. 3.52 during the combined A/B and B+C phases, respectively. Such results provide support for Hypotheses 3a and 3b.

Discussion

The present single-case dual-component multiple-baseline design and associated statistical analysis findings essentially replicate the mnemonic strategy "century" and "decade" results produced in earlier conventional "group" randomized experimental designs and analyses (Hwang & Levin, 2002; Hwang et al.,

1999). They also replicate portions of a recently reported, closely related single-case crossover design that focused on a mnemonic vocabulary strategy (Hwang et al., 2016). It is both comforting and encouraging to discover that effects that have been well established in the "group" intervention-strategy literature can be reproduced in comparably well-controlled small-sample (often drawn from low-incidence populations) research, such as in the present study based on only seven participants. This equivalence might ultimately prove to be a boon to intervention researchers working with limited resources, as reflected chiefly by the small number of available participants or the various "costs" associated with recruiting and testing multiple participants. We hasten to point out, however, that in comparison to conventional "group" randomized intervention research, single-case intervention research, with its much fewer number of participants, requires a more plentiful series of observations/sessions that furnish both baseline and intervention outcomes.

Our study's methodological and statistical approaches are applicable in any single-case context where an educational intervention researcher's focus is on assessing the efficacy of two different within-person intervention strategies, procedures, processes, and the like. Importantly, a well-crafted single-case multiple-baseline design possesses similar internal validity and scientific credibility characteristics to those of conventional "group" experimental or randomized controlled trials (RCT) intervention research studies (e.g., Levin, 1992; Levin et al., 2018), including, especially, studies in which various types of randomization are incorporated into the design and statistical analysis (Ferron & Levin, 2014; Kratochwill & Levin, 2010).

Can Single-Case Intervention Studies Serve as Attempted Replications of Conventional "Group" Design Findings?

Of specific relevance to the present study's replication goal, a reviewer of an earlier version of this article posed a provocative question, which perhaps was even issued as a challenge: "Would you hypothesize that single-case designs could replicate essentially any existing intervention effect?" Two initial clarifying responses are: First, single-case designs are not all-purpose research panaceas. Their primary research

objectives are associated with specific times and places (again see Horner & Odom, 2014). Second, and more important, we know full well that no intervention design—not even exactly the same design that was just implemented—can be counted on to replicate previous results (see, for example, Open Science Collaboration, 2012). So, given that the reviewer's intended question was whether there are single-case intervention designs currently available (or that could be constructed) that can *attempt* to replicate the findings produced by all (or almost all) conventional "group" intervention research investigations, we would reply with a cautious "More than likely," and here's why.

In the fields of psychological and educational experimental research there has been a long history of determining whether effects produced in between-group designs "stand up" (i.e., can be replicated) in analogous within-subject designs. What researchers have found is sometimes yes, sometimes no. The same speculations can be extended from conventional within-subject designs with their multiple participants and few outcome observations to analogous "within-subject" single-case (*aka* "interrupted time series") designs with few participants and multiple outcome observations. That said, with some ingenuity and selected design modifications to accommodate the resource constraints typically inherent in single-case intervention research, we would argue that it is possible to construct single-case design analogs of virtually all conventional "group" intervention designs for intended replication attempts. We now provide just a few examples to support that argument. For each of the single-case design analogs presented, we assume that case-replication is possible (i.e., that more than one case is included in the study) and that various types credibility-enhancing types of randomization have been incorporated into the design and statistical analysis (e.g., Ferron & Levin, 2014; Kratochwill & Levin, 2010).

The most straightforward example is the correspondence between conventional "group" pretest-intervention-posttest designs on the one hand and single-case AB-type and multiple-baseline designs on the

other, where A is a baseline or control phase with multiple measures (i.e., pretests) and B is an intervention phase with multiple measures (i.e., posttests). Such single-case designs can be constructed to examine the effects of a single intervention, as well as the combined and comparative effects of two or more interventions in multifactor and blocking designs (Levin et al., 2018; Levin & Wampold, 1999). Single-case alternating treatment designs and analyses can also be implemented to compare two or more different interventions (A, B, C, etc.) directly in a randomly alternating fashion (Gafurov & Levin, 2018; Horner & Odom, 2014; Kratochwill et Levin, 2010); and single-case "dual-order" and "crossover" designs can be fashioned to mimic those in the conventional "group" intervention researcher's toolkit (e.g., Hwang et al., 2016; Levin, Ferron, & Gafurov, 2014). As a final more complex example, single-case multilevel modeling designs and analytic approaches can be constructed to emulate those of conventional "group" designs with clustered units, such as classroom- or school-based interventions, including those that incorporate various covariate control factors (Rindskopf & Ferron, 2014). For now, such examples provide a limited response to the journal reviewer's "challenge."

Pragmatic Considerations in the Present Study

One particularly powerful form of single-case randomization is Edgington's (1975) concept of randomly determining the time point (here, the session) at which the intervention is introduced to each participant (e.g., Levin et al., 2018). In our study, intervention start-point randomization was not possible because of the relatively small number of total number of sessions that were available (15). That situation necessitated our choosing between a design based on multiple potential intervention start points with fewer participants and one based on more participants with a single fixed staggered intervention start point for each participant. Our ultimate decision was based on both statistical power (Levin et al., 2018) and pragmatic factors (e.g., time, student availability, and scheduling constraints), which led us to select the "more participants, fixed start points, alternative."

Similar pragmatic considerations resulted in our providing a between-student multiple-baseline stagger of only one session (rather that a preferred two- or three-session stagger). Quite simply, with seven students in the study and our decision to include at least three sessions devoted to each component (A, B, B+C), this was the only possible stagger option available. We are aware that the modified Revusky and Wampold-Worsham randomization-test procedures we applied are optimally sensitive to what are called "immediate abrupt" intervention effects (Levin, Ferron, & Gafurov, 2017). To produce such effects, the participant in each staggered tier must display a precipitous change in the level of the outcome measure *immediately* following the introduction of the intervention to that participant. On the negative side of that reality, if the intervention were to produce other than immediate abrupt effects, such as delayed abrupt effects or any type of gradual effects, the randomization-test analysis would experience a nontrivial loss of statistical power (Levin et al., 2017). However, on the positive side, given that each of the two mnemonic-invention components we introduced (century and decade in Stages 1 and 2, respectively) was expected to have a dramatic immediate impact on the levels of students' performance, we were cautiously optimistic that we could "get by" with only a one-session stagger – and fortunately, the results bore that out (again, please refer to Figure 1 for mnemonic century strategy effects and to Figure 4 for mnemonic decade strategy effects).

In addition, and reiterating Hwang et al.'s (2016) concerns, that there might be only two or three outcome observations (i.e., sessions) per phase in the present study would trouble an ardent visual analyst. Yet, a quantitative analyst would argue that each outcome observation encompassed four cognitive test items (rather than individual participant responses, observer ratings, behavioral observations, etc.), which should provide increased stability and confidence in those "individual" observations. Here, the visual analyst would arrive at a conclusion about the intervention's effectiveness primarily by examining the seven children's individual responses to the intervention, whereas the statistical analyst would rely primarily on an aggregated

measure based on between 5 and 14 of the children's outcomes (namely, a between-phase increase in the mean number of correct responses).

> Visual-analysis experts may disagree with statistical-analysis experts (and with one another) about whether a reliable intervention effect was produced on a case-by-case basis. By considering the pluses and minuses associated with the individual cases, visual analysts formulate a conclusion about intervention effectiveness in the study as a whole. *ExPRT* randomization tests, on the other hand, are focused on aggregated summary measures, which will lead the analyst to the same statistical conclusion every time. Even in situations where not all (or even none) of the individual case profiles would document an intervention effect through visual analysis, a statistical analysis based on the combined cases might. (Hwang et al., 2016, p. 11)[3]

Additional Considerations and Concerns in the Present Study

With respect to Figures 1, 3, and 4, some might argue that what was observed was simply a general increase in the children's performance level over the course of the study — a Shadish et al. (2002) "maturation"-type learning effect that really had nothing to do with the mnemonic interventions' efficacy. What must be kept in mind, however, is that even though an increase in level occurred between the study's first and second phase, the critical aspect of the Figure 3 data is that the initial increase in the children's level is generally not coincident with the staggered introduction of the mnemonic decade (B+C) intervention. In contrast, in Figures 1 and 4, the increases in the children's levels generally are coincident with the staggered introduction of the mnemonic century and mnemonic decade intervention (B and B+C, respectively). That "coincident" criterion is accounted for in the multiple-baseline randomization test that was conducted, which

is why the between-phase increase in level in Figure 3 is not statistically significant, whereas the between-phase level increases in Figures 1 and 4 are.

As was mentioned in the Results section and indicated in Table 3, only equivocal support for Hypothesis 2a's discriminant validity notions can be claimed for the Stage 2 century-memory results, where it can be argued that ceiling effects contributed to the statistically nonsignificant results that were observed there (Levin, 1985). Our attempt to produce a more "depressed" level of the children's century memory by focusing on their delayed-test performance was only partially successful. Even on the delayed tests, with only four three-choice (1700s, 1800s, 1900s) centuries for the children to remember, at- or near-ceiling level performance persisted.

Future Research Considerations

To afford more equitable future assessments of discriminant validity based on the present dual-component strategy approach, we recommend implementing the following modified and improved procedures:

1. Develop and incorporate two outcome measures that yield comparable score ranges, with no ceiling or floor effects. In addition, include more items to increase test difficulty.

2. Compare two strategies of similar complexity. In the present study, the mnemonic decade strategy was far more complicated, and more difficult to apply, than the mnemonic century strategy. With the present invention task, the mnemonic 10-level decade strategy (e.g., 17**70**s, 18**40**s) could comprise the first strategy component and a mnemonic 10-level "units" strategy (e.g., 177**6**, 184**8**) could comprise the second. A mnemonic unit strategy has already been developed and successfully implemented with older students (Hwang et al., 1999).

3. Include a multiple-baseline stagger for both component strategies. In the present study, the total number of children and sessions available permitted a stagger only for the first component strategy (i.e., the mnemonic century strategy). With more sessions, staggering both component strategies (century and decade) would have been possible, thereby producing a "cleaner" multiple-baseline assessment of discriminant-validity notions.

Even with such modifications and improvements, one must be mindful of the issues and complexities involved in marshalling convincing support for the "no effect" portion of the discriminant-validity argument, resulting from the myriad philosophical, methodological, psychometric, and statistical issues surrounding statistically nonsignificant findings as "proof" of the null hypothesis (e.g., Harlow, Mulaik, & Steiger, 1997; Serlin & Lapsley, 1999).

Finally, even carefully controlled and implemented single-case intervention studies might be faulted because of their limited population generalizability resulting from the small number of participants involved (an external validity consideration). From an internal validity standpoint, however, one would be hard pressed to pinpoint in those studies any flaws or plausible rival hypotheses that account for the results. Moreover, it can be argued that with appropriate randomization and control strategies built in, single-case multiple-baseline studies possess the same (or a highly similar) degree of internal validity as conventional randomized "group" studies (Kratochwill & Levin, 2010; Hwang et al., 2016). It is hoped that educational and psychological intervention researchers begin to recognize the methodological and statistical virtues of recently developed single-case intervention approaches in general (e.g., Kratochwill & Levin, 2010, 2014), and of multiple-baseline approaches such as the one presented here in particular.

References

Barton, E. E., & Reichow, B. (2012). Guidelines for graphing data with Microsoft® Office 2007™, Office 2010™, and Office **f**or Mac™ 2008 and 2011. *Journal of Early Intervention*, *34*, 129-150.

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis* (pp. 187–212). Hillsdale, NJ: Erlbaum.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81-105.

Dugard, P., File, P., & Todman, J. (2012) *Single-case and small*-n *experimental designs: A practical guide to randomization tests* (2nd ed.). New York: Routledge.

Edgington, E. S. (1975). Randomization tests for one-subject operant experiments. *Journal of Psychology*, *90*, 57-58.

Edgington, E. S., & Onghena, P. (2007). *Randomization tests* (4th ed.) Boca Raton, FL: Chapman & Hall/CRC.

Evans, J. J., Gast, D. L., Perdices, M., & Manolov, R. (Eds.) (2014). Single case experimental designs. Special issue of *Neuropsychological Rehabilitation: An International Journal*, *24*(3-4).

Ferron, J. M., & Levin, J. R. (2014). Single-case permutation and randomization statistical tests: Present status, promising new developments. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Statistical and methodological advances* (pp. 153-183). Washington, DC: American Psychological Association.

Ferron, J. M., Moeyaert, M., Van den Noortgate, W., & Beretvas, S. N. (2014). Estimating causal effects from multiple-baseline studies: Implications for design and analysis. *Psychological Methods*, *19*, 493-510.

Gafurov, B. S., & Levin, J.R. (2018, Dec.). *ExPRT* (*Excel Package of Randomization Tests*): *Statistical analyses of single-case intervention data* (Version 3.3). Downloadable from http://ex-prt.weebly.com/.

Harlow, L. L., Mulaik, S. A., & Steiger, J. A. (Eds.). (1997). *What if there were no significance tests*? Mahwah, NJ: Erlbaum.

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2013). A standardized mean difference effect size for multiple baseline designs across individuals. *Research Synthesis Methods*, *4*, 324-341.

Horner, R. H., & Odom, S. L. (2014). Constructing single-case research designs: Logic and options. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 27-51). Washington, DC: American Psychological Association.

Horner, R., & Spaulding, S. (2010). Single-case research designs. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 1386–1394). Thousand Oaks, CA: Sage.

Hwang, Y., & Levin, J. R. (2002). Examination of middle-school students' independent use of a complex mnemonic system. *Journal of Experimental Education*, *71*, 25-38.

Hwang, Y., Levin, J. R., & Johnson, E. W. (2018). Pictorial mnemonic-strategy interventions for children with special needs: Illustration of a multiply randomized single-case crossover design**.** *Developmental Neurorehabilitation*, *21*, 223-237. DOI: 10.3109/17518423.2015.1100689.

Hwang, Y., Renandya, W. A., Levin, J. R., Levin, M. E., Glasman, L. D., & Carney, R. N. (1999). A pictorial mnemonic numeric system for improving students' factual memory. *Journal of Mental Imagery*, *23*, 45-69.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, *2*, 696–701.

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education*, *34*, 26-38.

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods*, *15*, 122-144.

Kratochwill, T. R., & Levin, J. R. (Eds.). (2014). *Single-case intervention research: Methodological and statistical advances*. Washington, DC: American Psychological Association.

Levin, J. R. (1985). Some methodological and statistical "bugs" in research on children's learning. In M. Pressley & C. J. Brainerd (Eds.), *Cognitive learning and memory in children* (pp. 205-233). New York: Springer-Verlag.

Levin, J. R. (1992). Single-case research design and analysis: Comments and concerns. In T. R. Kratochwill & J. R. Levin (Eds.), Single-case research design and analysis: New developments for psychology and education (pp. 213-224). Hillsdale, NJ: Erlbaum.

Levin, J. R. (1994). Crafting educational intervention research that's both credible and creditable. *Educational Psychology Review*, *6*, 231-243.

Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2014). Improved randomization tests for a class of single-case intervention designs. *Journal of Modern Applied Statistical Methods*, *13*(2), Article 2; retrievable from http://digitalcommons.wayne.edu/jmasm/vol13/iss2/2. (Invited article)

Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2017). Additional comparisons of randomization-test procedures for single-case multiple-baseline designs: Alternative effect types. *Journal of School Psychology*, *63*, 13-34.

Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2018). Comparison of randomization-test procedures for single-case multiple-baseline designs. *Developmental Neurorehabilitation*, *21*, 290-311.

Levin, J. R., & Wampold, B. E. (1999). Generalized single-case randomization tests: Flexible analyses for a variety of situations. *School Psychology Quarterly*, *14*, 59-93.

Michiels, B., & Onghena, P. (2018). Randomized single-case AB phase designs: Prospects and pitfalls. *Behavior Research Methods*, Online version retrievable at https://doi.org/10.3758/s13428-018-1084-x.

Open Science Collaboration (2012). An open, large scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*, 657–660.

Parker, R. I., Vannest, K. J., & Davis, J. L. (2014). Non-overlap analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 127-151). Washington, DC: American Psychological Association.

Revusky, S. H. (1967). Some statistical treatments compatible with individual organism methodology. *Journal of the Experimental Analysis of Behavior*, *10*, 319-330.

Rindskopf, D. M., & Ferron, J. M. (2014). Using multilevel models to analyze single-case design data. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 221-246). Washington, DC: American Psychological Association.

Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 199-228). Hillsdale, NJ: Erlbaum.

Shadish, W. R. (Ed.). (2014). Analysis and meta-analysis of single-case designs. Special issue of the *Journal of School Psychology*, *52*.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Wampold, B., & Worsham, N. (1986). Randomization tests for multiple-baseline designs. *Behavioral Assessment*, *8*, 135-143.

Footnotes

We are grateful to Elizabeth Hampton, teacher at Zeeland Quest Elementary School, Fonda Green, director of the Children's After School Achievement (CASA) program, and Megan Swank, teacher at CASA, all of whom helped with student recruitment; to Hope College student Stacia Tibbetts for locating pictures of the vocabulary items; and to David James, Department of English, Hope College, for his invaluable assistance with the preparation of the manuscript. Correspondence regarding the study should be addressed to the authors at either hwang@hope.edu or jrlevin@u.arizona.edu.

1. For these hypotheses, we use the symbol "≈" to represent an equal or negligible amount.
2. A single analysis that examines the "differential effect" of the intervention on the X and Y measures (i.e., Levin & Wampold's, 1999, "comparative effect" hypothesis test) could also have been conducted but it does not provide evidence directly bearing on discriminant-validity specifications.
3. A visual analyst would be concerned about increasing baseline trends (especially, for example, that of Child 5) but that is not a relevant consideration in the "whole-series" modified Revusky and Wampold-Worsham randomization-test procedures that were applied here (see also, Michiels & Onghena, 2018).

Table 1. Illustration of a Traditional Multiple-Baseline Design, With Cases Randomly Assigned to the Four Tiers

	Time Period/Session													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Case 1	A	A	A	A	B	B	B	B	B	B	B	B	B	B
Case 2	A	A	A	A	A	A	B	B	B	B	B	B	B	B
Case 3	A	A	A	A	A	A	A	A	B	B	B	B	B	B
Case 4	A	A	A	A	A	A	A	A	A	A	B	B	B	B

Note: There is a single outcome measure (Y) that is administered to all cases in each session during both the baseline (A) and intervention (B) phases.

Table 2. Illustration of the Present Study's Sequentially Introduced Two-Component Multiple-Baseline Design, With Each Child Randomly Assigned to the Design's Seven Tiers

	Time Period/Session														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Child 1	A	A	A	B	B	B	B+C	B+C	B+C	B+C	B+C	B+C	B+C	B+C	B+C
Child 2	A	A	A	A	B	B	B	B+C	B+C	B+C	B+C	B+C	B+C	B+C	B+C
Child 3	A	A	A	A	A	B	B	B	B+C	B+C	B+C	B+C	B+C	B+C	B+C
Child 4	A	A	A	A	A	A	B	B	B	B+C	B+C	B+C	B+C	B+C	B+C
Child 5	A	A	A	A	A	A	A	B	B	B	B+C	B+C	B+C	B+C	B+C
Child 6	A	A	A	A	A	A	A	A	B	B	B	B+C	B+C	B+C	B+C
Child 7	A	A	A	A	A	A	A	A	A	B	B	B	B+C	B+C	B+C

Note: There are two outcome measures (X and Y) that are administered to all children during each session of the baseline (A), the first-stage intervention (B), and the second-stage intervention (B+C) phases. Measure X is hypothesized to be more sensitive than Measure Y to the effects of Intervention B and Measure Y is hypothesized to be more sensitive than Measure X to the effects of Intervention C.

Table 3. Discriminant Validity Hypotheses, Statistical Outcomes, and Hypothesis-Support Conclusions

Measure	Hypothesis	Statistical Outcomes		Hypothesis Support?
		p-value	Avg. Adjusted NAP	
Stage 1				
Delayed Century	H1a: B > A	.010	.74	Yes
Delayed Decade	H1b: B ≈ A	NS, A > B	.06 (A > B)	Yes
Vocabulary	H1c: B ≈ A	.455	.05	Yes
Stage 2				
Delayed Century	H2a: B+C ≈ A/B	.884	.39	Yes?
Decade	H2b: B+C > A/B	.001	.80	Yes
Vocabulary	H2c: B+C ≈ A/B	.234	.02	Yes

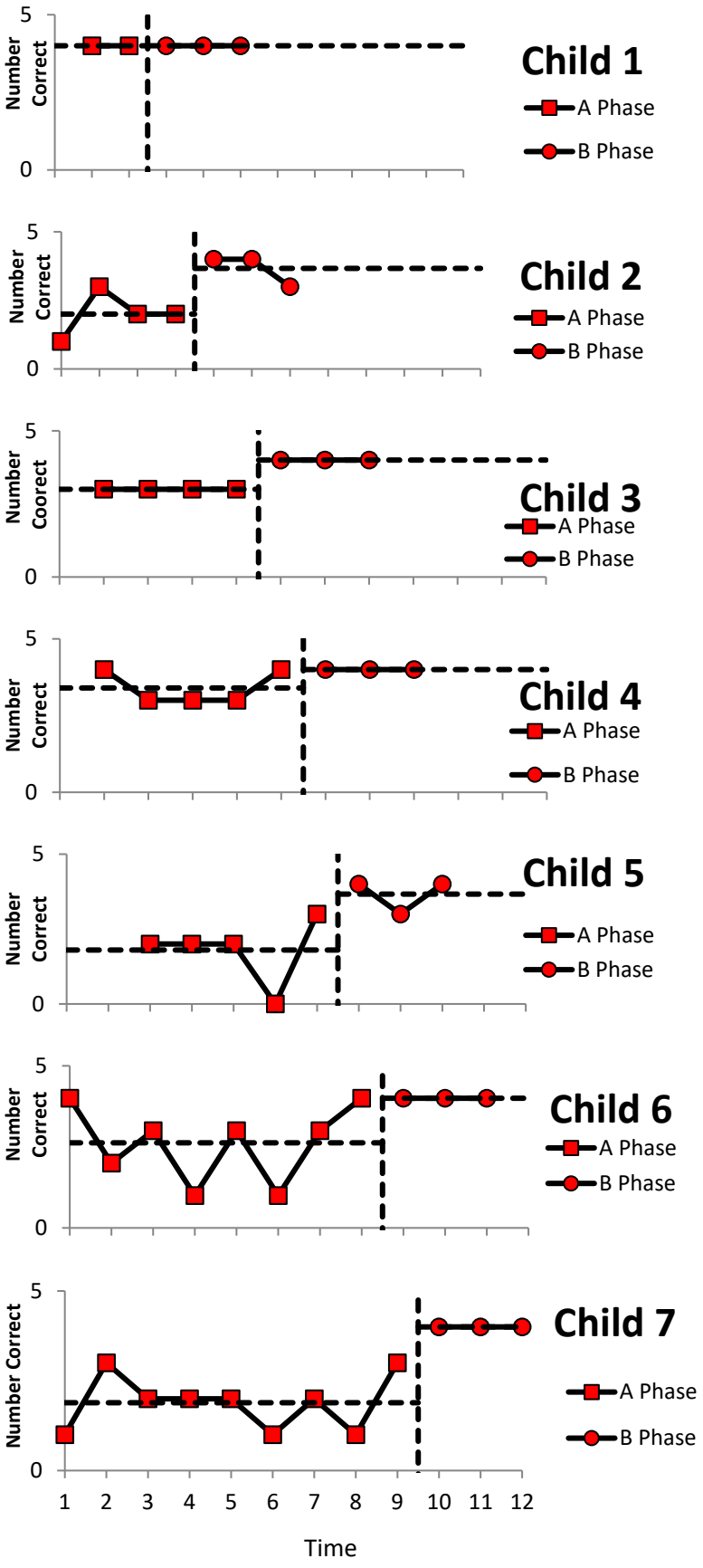
Figure Captions

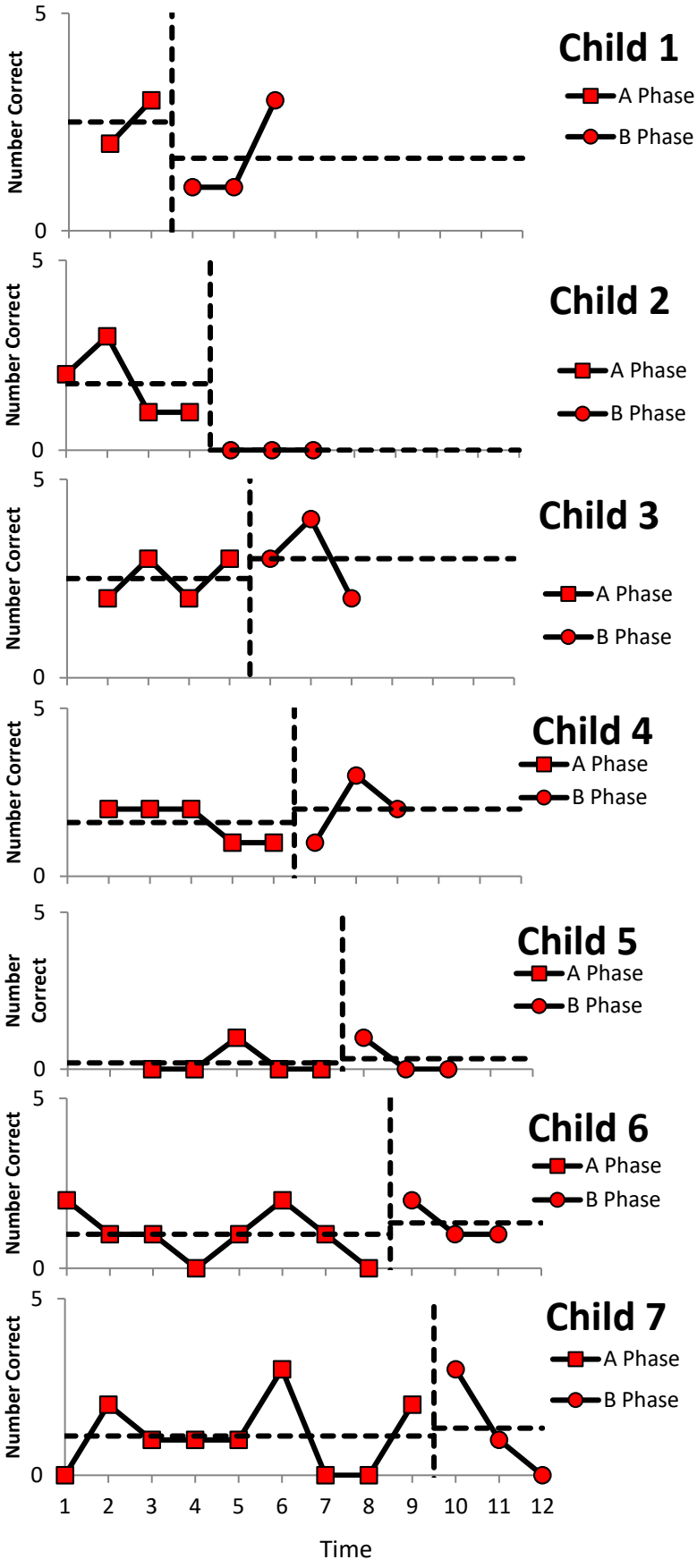
Figure 1. Stage 1: Number of Correct Responses on the Inventions Delayed Century Measure (A and B Phases)

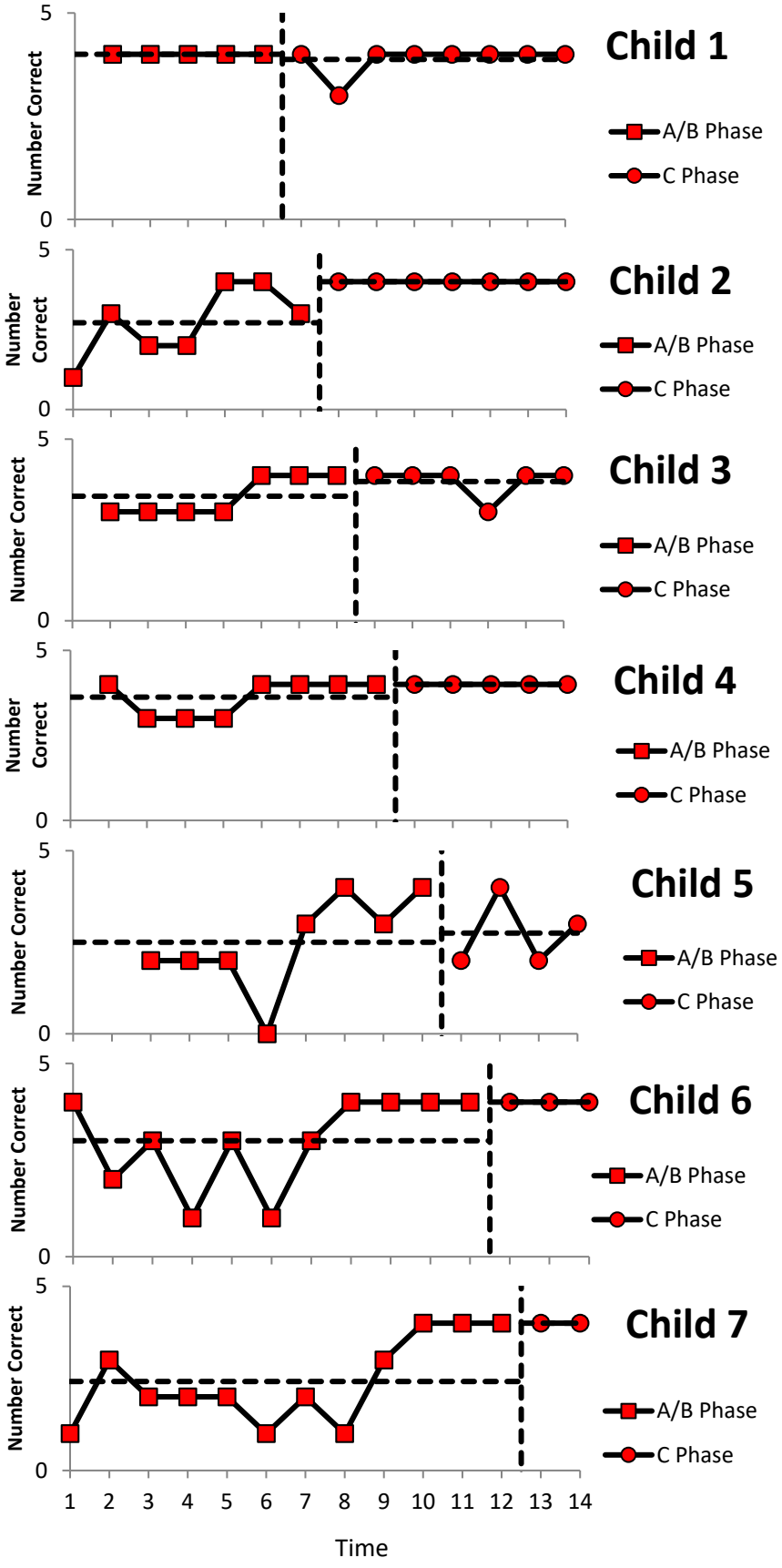
Figure 2. Stage 1: Number of Correct Responses on the Inventions Delayed Decade Measure (A and B Phases)

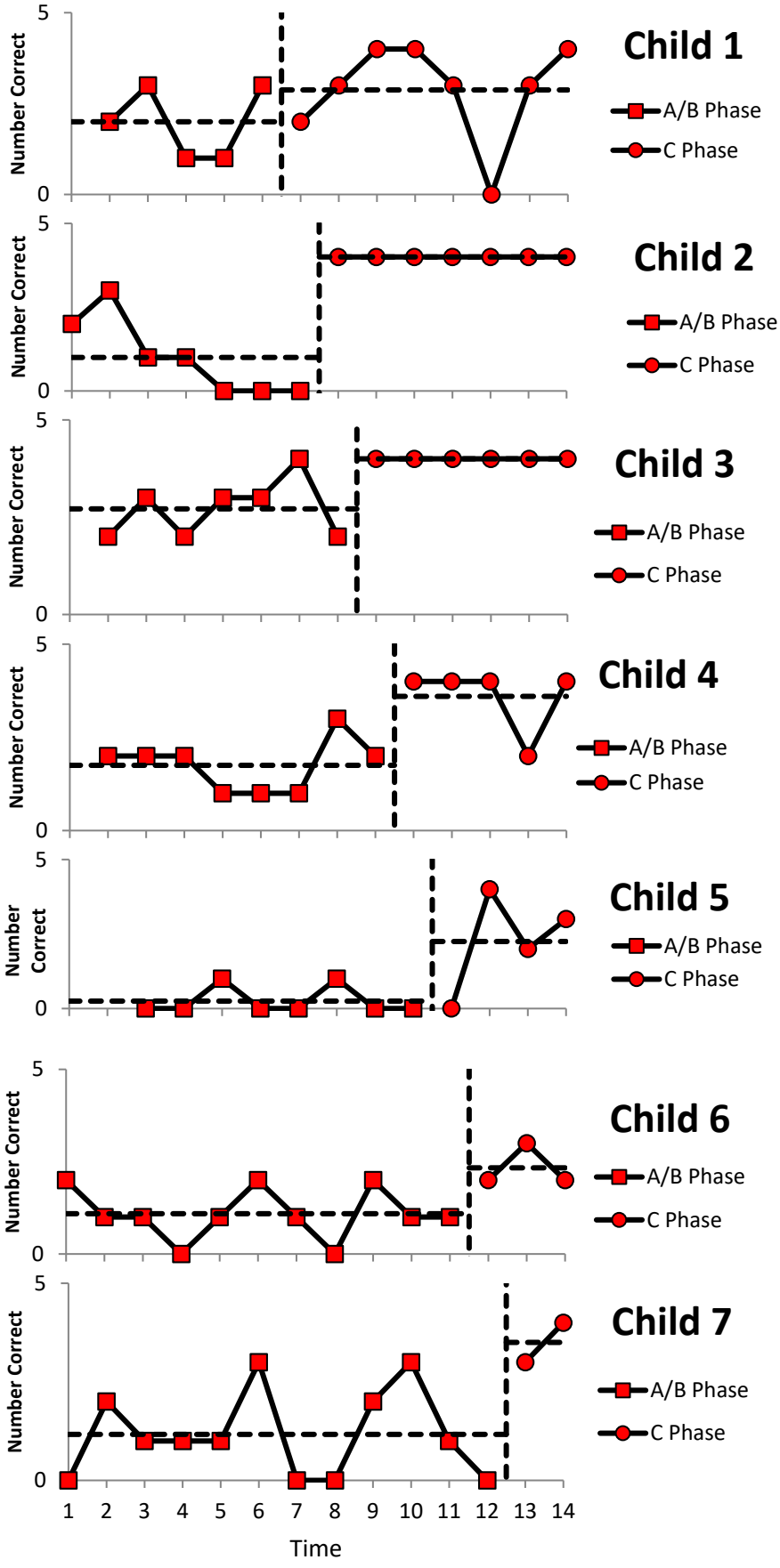
Figure 3. Stage 2: Number of Correct Responses on the Inventions Delayed Century Measure (A/B and C Phases)

Figure 4. Stage 2: Number of Correct Responses on the Inventions Delayed Decade Measure (A/B and C Phases)


Child 1
A Phase
B Phase
Number Correct
5
0
Child 2
A Phase
B Phase
Number Correct
5
0
Child 3
A Phase
B Phase
Number Coorect
5
0
Child 4
A Phase
B Phase
Number Correct
5
0
Child 5
A Phase
B Phase
Number Correct
5
0
Child 6
A Phase
B Phase
Number Correct
5
0
Child 7
A Phase
B Phase
Number Correct
5
0
1 2 3 4 5 6 7 8 9 10 11 12
Time


Child 1
A Phase
B Phase
Number Correct
5
0
Child 2
A Phase
B Phase
Number Correct
5
0
Child 3
A Phase
B Phase
Number Correct
5
0
Child 4
A Phase
B Phase
Number Correct
5
0
Child 5
A Phase
B Phase
Number Correct
5
0
Child 6
A Phase
B Phase
Number Correct
5
0
Child 7
A Phase
B Phase
Number Correct
5
0
1 2 3 4 5 6 7 8 9 10 11 12
Time


Child 1
Child 2
Child 3
Child 4
Child 5
Child 6
Child 7
Number Correct
A/B Phase
C Phase
0
5
1 2 3 4 5 6 7 8 9 10 11 12 13 14
Time


Child 1
Child 2
Child 3
Child 4
Child 5
Child 6
Child 7
A/B Phase
C Phase
Number Correct
Time
0
5
1 2 3 4 5 6 7 8 9 10 11 12 13 14