Demonstration of Single-Case Two-Sample Randomization-Test Procedures: Underperforming Children's Implementation of an Effective Memory Strategy

Yooyeun Hwang

Hope College

Joel R. Levin

University of Arizona

We are grateful to the following people for their help with student recruitment: Fonda Green, previous director of the Children's After School Achievement (CASA) program in Holland, Michigan; Terra Luckett, current director of CASA, and Kevin Hilgert and Rubi Solano, coordinators of CASA. We are also grateful to Hope College student Ayanna Bailey for her clerical support.

Abstract

Recently developed randomized-test procedures for two independent-samples single-case designs are presented and applied to a memory-strategy intervention study with eight underperforming students from low SES backgrounds. Research design aspects, data-analysis features, and various output measures are provided to demonstrate the potential utility of the randomization procedures for researchers who seek to examine comparative intervention effects with scarce resources but in a methodologically rigorous manner.

Demonstration of Single-Case Two-Sample Randomization-Test Procedures:

Underperforming Children's Implementation of an Effective Memory Strategy

A persistent challenge for researchers investigating academic interventions intended for students from low-performing populations is to conduct scientifically credible studies with "scarce resources" (Levin, 2007). Fortunately, however, in recent years a plethora of trustworthy single-case designs and associated data analyses, analogous to conventional large-sample intervention studies but based on only a handful of student participants, have been developed to fill that critical intervention-research niche (see, for example, Kratochwill et al., 2013; Tate et al., 2016). Such proposed single-case procedures have slowly but surely been gaining traction in both the methodological and experimental intervention research literatures.

Basic Single-Case Intervention Designs and Randomization Techniques

A basic single-case intervention design is an outgrowth of a conventional large-sample pretest-intervention-posttest design, except that it consists of one or more series of multiple pretest/control/baseline observations (A phases) followed by one or more series of multiple posttest/experimental/intervention observations (B phases) generally replicated across a few (say, 4 or 5) participants (e.g., Horner & Spaulding, 2010; Kazdin, 2021). Potential support for the desired intervention effect(s) can then be assessed both within participants (i.e., in their repeated ABAB…AB phase observations) and across the handful of participants in a systematically replicated AB design (e.g., Marascuilo & Busk, 1988). A variety of experimental controls must be incorporated into a single-case study's design and procedures to rule out potential internal invalidity threats, including but not limited to unwanted investigator and operational effects, participant effects (e.g., maturation, novelty, expectancy, order, testing effects), along with history and other confounding variable influences (Kratochwill et al., 2022; Shadish et al, 2002). Several randomization techniques have also been adopted to strengthen the effect inferences drawn from one-sample single-case intervention research, including inferences drawn from replicated AB designs, alternating treatment designs, multiple-baseline and multiple-probe designs, and changing criterion designs (see, for example, Craig & Fisher, 2019; Ferron & Levin, 2014; Ferron et al., 2019; Heyvaert & Onghena, 2014; Levin & Ferron, 2021; Levin et al., 2012, 2014, 2019b; Tanious & Manolov, 2022).

Two Independent-Samples Intervention Designs and Randomization Techniques

It is important to note that the conventional large-sample pretest-intervention-posttest design mentioned in the preceding paragraph does not even qualify as a *credible* intervention design. Campbell and Stanley (1966) refer to it as a pre-experimental design in that it does not adequately control for most of the potential confounding variables that were alluded to there. Adequate control would require a study based on two independently administered experimental conditions to which participants are randomly assigned, namely, a pretest-intervention-posttest condition and a pretest-placebo/alternative intervention-posttest condition. Although less stringent, more suitable, credibility criteria have been applied to replicated and carefully administered one-sample single-case intervention designs (Kratochwill et al., 2013) – and particularly to randomized multiple-baseline designs (Levin & Ferron, 2021) – there are situations in which actual single-case intervention research investigations would benefit from the adoption of randomized two independent-samples designs and data analyses. Such proposed models have recently appeared in the literature, for both replicated AB designs (Levin et al., 2019a) and for multiple-baseline designs (Levin et al., 2022). The major advantage of these single-case two-sample randomized designs is that following random assignment of participants to experimental conditions, there is a well-controlled between-conditions comparison of the focal intervention with a placebo or alternative intervention. The present study, based on a small sample of underperforming students drawn from a low-SES population, demonstrates how a randomized single-case two-sample replicated AB design and data analysis can constitute a single-case analog of a conventional large-sample randomized design in comparing students' acquisition of two different strategies for remembering the dates of various inventions (Hwang & Levin, 2002; Hwang et al., 2018, 1999).

Method

Participants

Eight fifth-grade students (average age, 10 years, 6 months; six girls, two boys) who attended a community after-school program that promoted educational achievement among underrepresented children through one-on-one tutoring participated in the study. The program accepts students who are struggling in their academics at school, as determined on referrals from their teachers. Most of the students in the program are from low-income households and qualify for free or reduced lunch (e.g., 83% in 2021-22). The eight students' primary language was English, but they were all bilingual. Specifically, seven of them could speak Spanish, and one spoke Swahili. Student participation consent forms were sent to, and approved by, the students' parents. The forms explained the study, the investigator's contact information, and the options to opt out anytime during the study process.

Research Design

A randomized two independent-samples single-case design and randomization-test analysis, as detailed by Levin et al. (2019a), allowed for the 8 participating students to be randomly assigned in equal numbers to two different invention-date mnemonic (memory-enhancing) strategy conditions, century only and century + decade. Each participant was randomly assigned either 3 baseline/control sessions and 4 intervention sessions or 4 and 3 sessions, respectively (for reasons explained later), with the intervention sessions featuring the mnemonic strategy condition to which the participant had been assigned. Unfortunately, after 6 students had completed their participation in the study, the research was interrupted by the Covid pandemic. Two students (one in each of the study's conditions) who had only partially completed the study because of weather and health conditions, had to be replaced after the program was resumed.

Study Materials

Fifty-six inventions were selected from three centuries (18 18th-century, 19 19th-century, and 19 20th-century) and were randomly grouped into seven lists of eight, with each list including no more than three inventions from the same century and none of the eight dates representing the same decade. Nine additional inventions were selected and used as examples in the initial session. Information about each invention was presented on 8.5 x 11-inch sheets of paper that included a line drawing of the invention name and its date consisting of a century and a decade; and for the mnemonic century and decade line drawings, additional illustrated characters, printed descriptions, and in some cases, captions associated with the characters. The presentation and testing orders of the eight items were individually randomized for each participating student in each session, with no sequence the same for any two students.

In the century mnemonic condition, following the baseline phase students were shown how to use a simple mnemonic numeric strategy to help them remember the century in which each presented item was invented. Specifically, the strategy consisted of illustrations of people from three different centuries (kings and queens for the 1700s, cowboys for the 1800s, and astronauts for the 1900s). Two example inventions were then presented in century settings. The experimenter reminded students that they could use whatever method they preferred to remember the decade associated with each of the inventions.

The same century-remembering strategy was implemented during the intervention phase of the century + decade mnemonic condition. A more complex mnemonic numeric strategy was added, which incorporated 10 months of the year to represent the decades (December for the 00s decade, January for the 10s decade, February for the 20s, etc., through September for the 90s). To make things more concrete and picturable, each month was associated with a prototypical seasonal setting [December (00s decade) = a Christmas scene; January (10s decade) = a New Year's Eve celebration; February (20s decade) = Valentine's Day…September (90s decade) = returning to school] in which the century characters (royalty, cowboys, or astronauts) were inserted and interacting. Again, two examples showed the students how to apply the methods to remember the inventions and their dates by combining the two techniques.

Procedures

Journal space requirements do not permit a complete account of the present procedures, and so only an abbreviated version of them is now provided. A more detailed description of the procedures for a related one-sample single-case intervention study can be found in a recent article by Hwang and Levin (2019).

For all students, in their initial baseline-phase session, the experimenter informed the students of the upcoming inventions-and-dates task, reminded the students what centuries and decades were in number of years, presented three example inventions (e.g., microwave, sea clock, wrench), informed the students that they would be shown eight inventions and dates from three centuries (1700s, 1800s, and 1900s), and that later they would take a test to see if they could remember both the century and decade of each invention. The experimenter then presented the eight inventions and dates in all students' initial baseline-phase session. First, she read the name and the dates (century and decade) to the student and pointed out the representational picture of the invention (see Panel a of Figure 1 for the 1950s invention of a milk carton), which the students studied in a self-paced fashion.

[INSERT FIGURE 1 ABOUT HERE]

Then, to mitigate short-term memory effects, after viewing all eight items the students completed a filler task in which they searched for 10 different words on a word-search sheet. An immediate invention-date recall test of the eight items was then administered, and in subsequent sessions this was followed by a delayed invention-date recall test of the eight items from the previous session. On these tests, students could either name the century and the decade or point and choose from the three centuries and the 10 decades written underneath the sentence. Because of absences, program cancelations, snow days, and holidays, the duration of the delay varied from two days to two weeks. Following each delayed test, the experimenter presented a new list of eight inventions, followed by the filler task and the immediate recall test.

In intervention-phase sessions following the 3 or 4 sessions of their assigned baseline phase, students in the two experimental conditions learned their specific mnemonic strategy to remember the dates. In the century mnemonic condition, the experimenter read the invention name and date to the students and then read the century setting (e.g., astronauts for the 1900s), along with the printed description and any caption in the interactive picture (see Panel b of Figure 1 for the 1950s milk carton invention). In the century + decade mnemonic condition, the experimenter read the invention name and date and then read the century and seasonal settings (e.g., astronauts for the 1900s and a flower garden for the 50s decade), along with the printed description and any caption in the interactive line drawing (see Panel c of Figure 1).

Results

As was mentioned earlier, one of two potential baseline-to-intervention transition points was randomly assigned to each participant, in accord with Koehler and Levin's (1988) randomization-test procedure, to increase the power of the ensuing statistical tests (Levin et al., 2019a). By chance, the transition point for all four century strategy students was between Session 3 and Session 4, whereas for the century + decade strategy, two students were randomly assigned a transition point between Session 3 and Session 4, and two between Session 4 and Session 5. The results, based on randomization tests with Type I error probabilities of .05 and analyzed through Gafurov and Levin's (2021) freely available *ExPRT* single-case randomization-test package, are now presented for the various outcome measures examined. Readers unfamiliar with the *ExPRT* package should refer to *ExPRT's* User Instructions for detailed information on how to set up and conduct the following randomization-test analyses.

[INSERT FIGURE 2 ABOUT HERE]

Immediate Recall Results

Century recall. Results for all student participants are presented in Figure 2. For the immediate century recall measure, it was predicted that there would be an abrupt between-phase performance increase in mean century recall in both experimental conditions immediately following the introduction of the century mnemonic strategy.

In *ExPRT's* (Version 4.2) AB design module, a randomization-test analysis of the "general" average between-phase change (Levin et al., 2019a) begins by entering on the Data sheet the eight students' immediate century scores, one row at a time for the four students in one of the conditions, followed by the four students in the other condition. Then, on the Interventions sheet, in Columns A through E are listed, one row at time, each of the student's particulars regarding the potential and actual intervention start points. For the present example, each student was randomly assigned one of two potential start points, beginning in either Session 4 or Session 5.

A one-tailed, $\alpha = .05$, test is specified in Columns G and H, with the expected baseline-to-intervention phase performance increase designated with a 2 in Column I. The remaining Interventions sheet input is straightforward, other than leaving the "Two-Group" specification in Column U empty for this "general intervention effect" analysis and being sure to specify an AB actual intervention order in Column T for each student. A click on the Run button reveals that the average across-student improvement in century recall was statistically significant, $p = .008$. Then, a click on the Plot button produces graphs and selected summary measures.

It was also predicted that the between-phase increases in century recall would be comparable in the two mnemonic strategy conditions, century and century + decade. The only changes required on *ExPRT's* Interventions sheet to conduct a two independent-samples randomization test of this "comparative intervention effect" are: (1) the test would likely be conducted with a two-tailed α of .05, as specified in Columns G and H; (2) a 3 or 4 would be specified in Column I, depending on which condition's four student scores were listed first on the Data sheet; and (3) a 4 would be specified in Column U to indicate that a two-group comparison is desired.

A click on the Run button reveals that the comparative change in mean recall – between the baseline (A) phase and the intervention (B) phase in the two mnemonic strategy conditions – was not statistically significant, $p > .50$, with an average standardized improvement difference between conditions of .41 favoring the mnemonic century condition. The average between-phase increase amounts to an average of 3.88 items in the century strategy condition (left side of Figure 2) and 3.23 items in the century + decade strategy (right side of Figure 2), which represents ceiling-approaching performance in both conditions. A reasonable explanation for why the former students' descriptively larger increase is somewhat higher than the latter's is that the former students' attention was directed more to (and assisted with) the century information than the decade information, whereas the latter students' attention and assistance was focused on both types of information.

[INSERT FIGURE 3 ABOUT HERE]

Decade recall. Results for all student participants are presented in Figure 3. For the immediate decade recall measure, it was predicted that in contrast to the immediate century recall results, the between-phase performance increase in the century + decade mnemonic condition would by far surpass that in the century mnemonic condition. Consistent with that prediction, here it may be seen that there are sizable differences between the two mnemonic strategy conditions in students' baseline phase to intervention phase decade-recall improvement, clearly favoring the mnemonic century + decade strategy condition (right side of Figure 3) over the mnemonic century condition (left side of Figure 3). Application of the same *ExPRT* randomization-test procedures as described above, but with the immediate decade-recall outcome measure, yielded the following results: (1) The across-conditions general intervention effect was statistically significant, $p = .01$, and as may be seen in Figure 3, this increase is attributable exclusively to the performance of the four students in the century + decade mnemonic condition; and (2) A randomization test of the performance difference between the two conditions (i.e., the comparative intervention effect) was statistically significant, yielding a *p*-value less than .001. All four students in the mnemonic century + decade strategy condition exhibited a substantial jump in their decade recall that was coincident with the introduction of their strategy instruction (by an average of 4.65 items, an increase of 279%), as compared to the decade recall in the mnemonic century strategy condition (an average improvement of .38 items, an increase of 22%). The difference between the average increases in the two conditions represents 4.41 standard deviation units. In the mnemonic century condition, only one student's (Student 4's) decade recall showed some improvement, and that just in the final intervention session. Parker and Vannest's (2009) between-phase data nonoverlap NAP index, rescaled to range between 0 and 1, was .23 in the mnemonic century condition, as compared to .98 (virtually no between-phase overlap in outcomes) in the mnemonic century + decade condition.

Century + decade recall. Because these results essentially mimic those for the decade recall measure, no Figure is presented for them. Specifically, the two-sample randomized test *p*-value, favoring the mnemonic century + decade condition, was less than .001 and the mean between-phase improvement difference was 4.00 recalled century-plus-decade items, a standardized difference of 3.81 units. The average between-phase NAP for mnemonic century strategy students was .63, compared to 1.00 (complete nonoverlap between phase outcomes) for mnemonic century + decade students.

Delayed Recall Results

The delayed recall findings are potentially interesting but because of different, and uncontrolled, amounts of between-session time delays within and between students, between baseline and intervention sessions, and especially between the two intervention conditions, those findings are not reported in detail here. In short, students in the mnemonic century + decade condition outperformed their counterparts in the mnemonic century condition on both the delayed century recall and delayed decade recall measures. However, the former students experienced a lesser average amount of between-session delay during the intervention phase (and therefore less time for forgetting information from the previous session) as compared to the latter students, thereby making the delayed recall measures difficult to interpret.

Concluding Comments

The present study illustrates (both figuratively and literally) how an intervention researcher working with very small samples and often targeting low-performing students, can compare the effectiveness of two different interventions (or an intervention with either a placebo control or a “business as usual” condition) through a recently a developed single-case two independent-samples randomized design and randomization-test procedure. In our example, students who were taught to employ a highly effective complex mnemonic strategy by far outrecalled students who applied a simple mnemonic strategy and adopted their own preferred remembering strategy on a task that favored use of the complex mnemonic strategy. It is important to note that the present complex mnemonic strategy differentially facilitated the students’ performance specifically on the outcome measure for which it was designed, namely decade recall. As has been well established in the learning-and-memory literature, this goes to show that effective mnemonic strategies are not universal panaceas that improve students’ performance across the board, but rather only the specific substantive content that the mnemonic components were designed to target (see, for example, Hwang & Levin, 2019; and Hwang et al., 2018). At the same time, the present underperforming students in both mnemonic conditions reported having enjoyed their learning experience.

A within-case randomization model for comparing two alternative interventions (or a baseline and intervention condition) might be considered in a one-sample replicated AB or multiple-baseline format as well, as has recently been demonstrated by Hwang and Levin (2019). In those designs, however, researchers must be careful to consider carryover effects, order, testing, and maturation effects, comparison effects, novelty effects, reactivity, and the like – all of which are well controlled in a randomized two independent-samples single-case design. The present replicated AB two-sample model, with its random assignment of participants to the two experimental conditions and its random assignment of intervention start points to participants, represents a methodologically rigorous single-case intervention design. With the addition of the associated randomization test, the model has been found to provide a reasonably powerful statistical procedure for detecting comparative intervention-effect differences of moderate size (Levin et al., 2019a).

Two operational issues are worth mentioning. First, in the traditional single-case behavior analysis literature, the outcome measures often consist of a participant's discrete (e.g., 0-1) behavioral responses or of an interventionist's observational ratings of the participant's performance. Such measures are unlikely to be as reliable or as stable as the eight-item memory tests that were administered in the present study (see, for example, the present Figure 3). With less reliable measures, a greater number of outcome observations would be required to produce accurate assessments of the participants' performance than the 7-observation series that was implemented here (Levin et al., 2019a; see also Hwang et al., 2018, p.233). Second, developing *a priori* unbiased rules for dealing with participant dropouts is a challenge that faces all single-case interventionists and should not be ignored.

That said, intervention researchers conducting "scarce resources intervention studies" are encouraged to consider the newly developed two-sample randomization procedures even when expecting less dramatic comparative-intervention differences than were observed here, but still in the context of a scientifically convincing investigation.

References

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Rand McNally.

Craig, A. R., & Fisher, W. W. (2019). Randomization tests as alternative analysis methods for behavior analytic data. *Journal of the Experimental Analysis of Behavior*, *111*, 309-328. https://doi.org/10.1002/jeab.500

Ferron, J. M., Rohrer, L. L., & Levin, J. R. (2019). Randomization procedures for changing criterion designs. *Behavior Modification*. Retrievable at https://doi.org/10.1177/0145445519847627.

Gafurov, B. S., & Levin, J. R. (2021, November). *ExPRT* (*Excel Package of Randomization Tests): Statistical Analyses of Single-Case Intervention Data*; current Version 4.2.1 is retrievable from the *ExPRT* website at http://ex-prt.weebly.com

Heyvaert, M., & Onghena, P. (2014). Randomization tests for single-case experiments: State of the art, state of the science, and state of the application. *Journal of Contextual Behavioral Science*, *3*(1), 51–64. https://doi.org/10.1016/j.jcbs.2013.10.002

Horner, R., & Spaulding, S. (2010). Single-case research designs (pp. 1386–1394). In N. J. Salkind (Ed.), *Encyclopedia of Research Design*. 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. (2019). Application of a single-case intervention procedure to assess the replicability of a two-component instructional strategy. *Contemporary Educational Psychology*, *56*, 161-170.

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.

Kazdin, A. E. (2021). *Single-case research designs: Methods for clinical and applied settings*. (3rd ed.). Oxford University Press.

Koehler, M. J., & Levin, J. R. (1998). Regulated randomization: A potentially sharper analytical tool for the multiple-baseline design. *Psychological Methods*, *3*, 206-217.

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.

Levin, J. R. (2007). Randomization tests: Statistical tools for assessing the effects of educational interventions when resources are scarce. In S. Sawilowsky (Ed.), *Real data analysis* (pp. 115-123). Greenwich, CT: Information Age.

Levin, J. R., & Ferron, J. M. (2021). Different randomized multiple-baseline models for different situations: A practical guide for single-case intervention researchers. *Journal of School Psychology*, *86*, 169-177. https://doi.org/10.1016/j.jsp.2021.03.003

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), 2-52; retrievable from http://digitalcommons.wayne.edu/jmasm/vol13/iss2/2.

Levin, J. R., Ferron, J. M., & Gafurov, B. S, (2019a). An improved two independent-samples randomization test for single-case AB-type intervention designs: A 20-year journey. *Journal of Modern Applied Statistical Methods*, *18*(1), Article 23, 1-20. Online version available at doi:10.22237/jmasm/15566/70480.

Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2022). Novel randomization tests for two-sample multiple-baseline designs. *Journal of Education for Students Placed at Risk*. Online version available at doi: 10.1080/10824669.2022.2045995.

Levin, J. R., Ferron, J. M., & Kratochwill, T. R. (2012). Nonparametric statistical tests for single-case systematic and randomized ABAB...AB and alternating treatment intervention designs: New developments, new directions. *Journal of School Psychology*, *50*, 599-624.

Levin, J. R., Kratochwill, T. R., & Ferron, J. M. (2019b). Randomization procedures in single-case intervention research contexts: (Some of) "The rest of the story". *Journal of the Experimental Analysis of Behavior*, *112*, 334-348.

Parker, R. I., & Vannest, K. J. (2009). An improved effect size for single case research: Non-overlap of all pairs (NAP). *Behavior Therapy*, *40*, 357-367.

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

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

Tanious, R., & Manolov, R. (2022). A practitioner's guide to conducting and analysing embedded randomized single-case experimental designs. *Neuropsychological Rehabilitation*. Advance online publication. https://doi.org/10.1080/09602011.2022.2035774

Tate, R. L., Perdices, M., Rosenkoetter, U., McDonald, S., Togher, L., Shadish, W., Horner, R., Kratochwill, T., Barlow, D. H., Kazdin, A., Sampson, M., Shamseer, L., & Vohra, S. (2016). The single-case reporting guideline in BEhavioural interventions (SCRIBE) 2016: Explanation and elaboration. *Archives of Scientific Psychology, 4*(1), 10-31. https://doi.org/10.1037/arc0000027

Figure 1. Sample instructional materials for teaching that the milk carton was invented in the 1950s: Panel a = baseline picture for all students; Panel b = intervention phase picture for students in the century mnemonic condition (with the 1900s century represented by astronauts); Panel c = intervention phase picture for students in the century + decade mnemonic condition (with the 1900s century represented by astronauts and the 50s decade represented by the fifth month, May, with flowers in a garden setting)

a b c

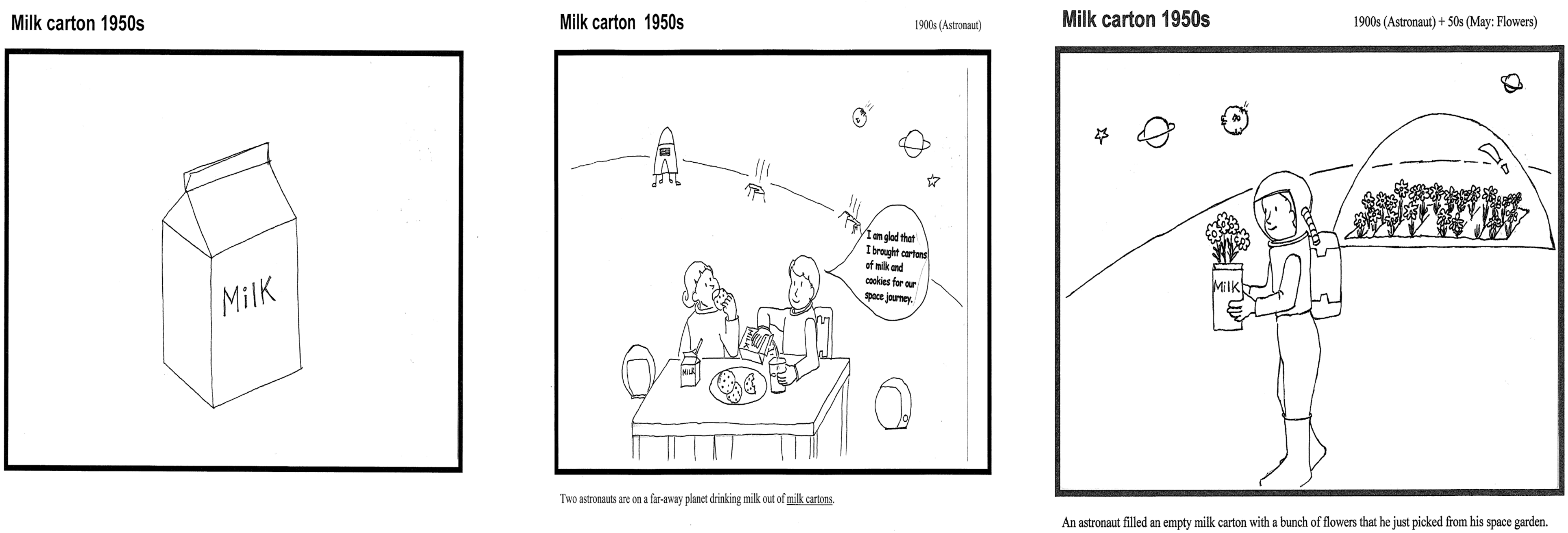

Figure 2. Number of correct century responses on the immediate test for students in the two instructional conditions

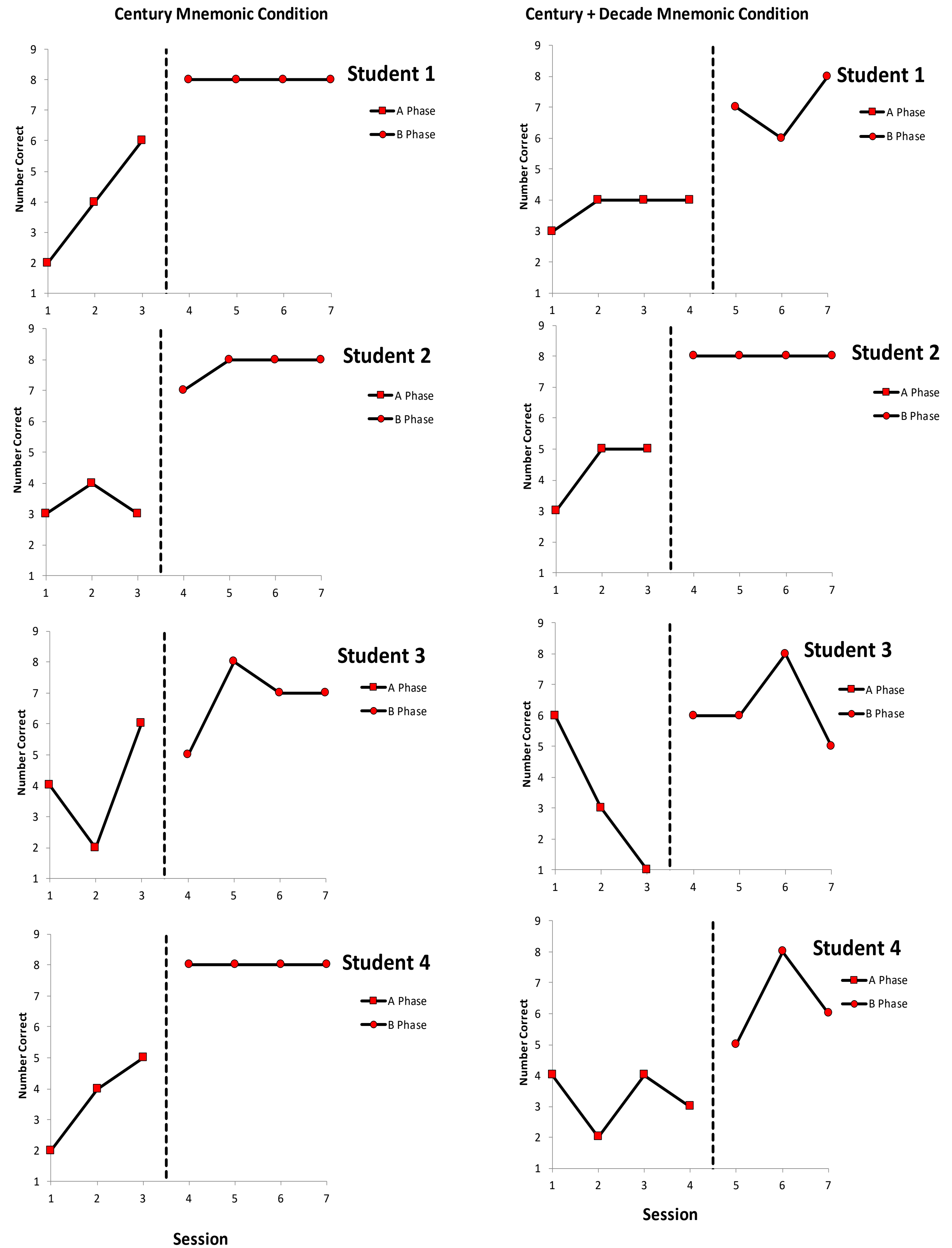

Figure 3. Number of correct decade responses on the immediate test for students in the two instructional conditions