Title
Systematic Replication Study of Interleaved Math Practice

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors
Matlen, Bryan
Bartel, Anna
Barrett, Drew
et al.

Publication Date
2026

Systematic Replication Study of Interleaved Math Practice

Bryan J. Matlen (bmatlen@wested.org)[1], Anna Bartel[1], Drew Barrett[1], Alexander Jacobson[1], Doug Rohrer[2], & Jodi Davenport[1]
[1]Center for Teaching & Learning, WestEd
[2]Department of Psychology, University of South Florida

Abstract

Interleaved practice involves mixing problems of different types within assignments and distributing problems of the same type over time. A recent classroom randomized study conducted by Rohrer et al. (2020) reported large learning effects of interleaved practice relative to blocked practice. In the present study, we test whether we can replicate the effects of Rohrer et al. (2020) while also systematically extending the interleaved intervention to enhance its suitability for educational contexts. We report on results of two of three planned cohorts (n=813 middle-school students across 8 districts), and find that the interleaved condition outperforms a blocked condition (g=.18, p<.05). Implementation analysis and a cost analysis suggest that the intervention can be implemented with fidelity at low cost to schools. We discuss implications of these findings as well as the broad challenge of translating learning research to practice.

Keywords: Interleaving, Math Education, Learning Research

Introduction

Consider a typical mathematics assignment: students complete a lesson calculating the volume of a prism, then are given a worksheet fifteen problems that all require applying the formula for prism volume. The following day, they learn about the area of rectangles and complete another fifteen problems on rectangles. This instructional approach, known as blocked practice, dominates mathematics curricula. Blocked practice follows a straightforward formula: introduce a concept, provide repeated practice with that concept. The practice is intuitive, there is a clear focus, and students can gain confidence as they solve problem after problem using the same approach.

Yet blocked practice has a critical flaw. It creates an illusion of mastery that is shattered when students must apply their knowledge flexibly, such as on a high stakes test. On homework assignments, students can correctly solve the problems without practicing an important skill they will need for later application - choosing which strategy to apply. Instead, in blocked practice, students can simply execute a provided procedure that is top of mind after a recent lesson. But when these same students encounter isomorphic problems three months later on a high stakes test their performance often deteriorates. They struggle because unlike during practice, they are not told which strategy to apply, and they cannot discriminate between problem types that require different procedures. Blocked practice, while beneficial for learning within problem similarities (e.g., Carvalho & Goldstone, 2014), fails on its own to develop the adaptive expertise that characterizes deep mathematical understanding.

In contrast, spaced interleaved practice (referred to as "interleaved practice" for brevity) involves rearranging problems in two ways: 1) assignments consist of problems of different types that are mixed together such that the same problems rarely follow one another during study (e.g., a problem requiring the pythagorean theorem is followed by calculating the area of a rectangle) and 2) the same problems are distributed across assignments so that students are exposed to them over time. This seemingly simple rearrangement of problems takes advantage of two powerful learning mechanisms: 1) discrimination learning - mixing different problems forces students to actively select relevant strategies based on problem features, improving discrimination abilities (Kang & Pashler, 2012), and 2) spaced practice - distributing problems of the same type across study sessions, requiring students to recall strategies over time, strengthening long-term retention (Lyle et al. 2020).

Studies consistently demonstrate that interleaved practice yields higher learning outcomes than blocked practice, particularly in mathematics (e.g., see Rohrer & Hartwig, 2020 for review). Moreover, the impact of interleaving was recently examined in a cluster randomized trial, comparing the effects of interleaving vs blocking in middle school mathematics classes (Rohrer, Dedrick, Hartwig, & Cheung 2020). The study took place in 54 seventh grade classrooms in a single district. Each class was randomly assigned within teacher to either the blocked or interleaved group, and the study took place over the course of a half year, receiving an assignment approximately once a week over that time. The results showed that the interleaved group outscored the blocked group by a large, statistically significant margin on a post-test administered 4 weeks after the intervention (Cohen's d = 0.83).

Despite this strong evidence; however, interleaved practice remains conspicuously absent from mainstream educational materials. As an illustration, Rohrer, Dedrick, and Hartwig (2020) examined six popular middle school mathematics textbooks for the presence of interleaved and blocked problem sets. They found that blocked practiced predominated each of the texts, and the number of interleaved problems in each text translated to an average of only one or two problems per school day.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

Translating Cognition Research Into Practice

The mismatch in research on interleaving and the sparsity of interleaving in practice resembles the paradox of other learning principles: Despite the natural link between learning research and education, learning research is not widely used in authentic learning contexts (Pomerance, Greenberg, & Walsh, 2016). One reason for this disconnect could be that learning takes place within complex systems that are highly contextual (Stigler et al. 2020). Translating learning research requires adapting principles to match the needs and constraints of specific contexts.

As an illustration, one key ingredient of interleaving involves mixing problems of different types within assignments. However, this general recommendation leaves open many unanswered instructional questions such as the number of items that are optimal within an assignment, the way in which the assignments should be administered (e.g., in class or as homework, independently or working in pairs), and so on. In the absence of such guidance, educators must make many 'best guess' decisions not directly informed by the past research on interleaving.

Another possible reason for the lack of adoption of interleaving in practice, despite strong classroom evidence from Rohrer et al. (2020), is that the design of this study had several limitations that may reduce its relevance to education stakeholders. For example, the study was conducted in a relatively isolated geographic region (i.e., a single school district in one state) and in a relatively homogeneous sample (excluding students not at proficiency based on the state test).

The Present Study

In part to address the shortcomings described above, we are in the process of conducting a systematic replication study to replicate and systematically extend the study conducted by Rohrer et al. (2020). The aim of the study is to test whether we can effectively scale the intervention to better match the constraints of educational contexts. Towards the aim of replication, we hold several features constant including the intervention and control conditions, classroom random assignment within teachers, and the age group and domain (middle-school mathematics). However, to enhance the generality of the findings for classroom contexts, we systematically vary several features including the sample - targeting a broader sample across multiple districts and states and including students regardless of prior ability - the timeperiod - extending the intervention over a school year - and the intervention delivery - we use digital as opposed to paper materials through ASSISTments (Heffernan & Heffernan, 2014). The study is conducted independently by WestEd, and Dr. Rohrer consults on the design and materials to ensure consistency with the past study. The research questions were the following:

1. *Impact.* What is the impact of interleaved practice relative to blocked practice on middle-school students math achievement, measured 4-weeks after the intervention? (test of replication)
2. *Fidelity.* To what extent can students and teachers implement interleaving with fidelity?
3. *Cost.* What is the cost of the interleaved intervention relative to the blocked condition and a business-as-usual case?

Towards these aims we report on the findings of the first two of three planned cohorts from this study. In addition to examining the efficacy of interleaving on student learning we report on fidelity of the implementation and present a cost analysis of interleaved practice relative to blocked practice and a business-as-usual control.

To recap, our proposed study differs from Rohrer et al. (2020) in that we plan to conduct our study in a more heterogeneous sample and broader geographic region. In education research, larger and more heterogeneous samples often result in smaller observed effects (e.g., Cheung & Slavin, 2016); thus, we anticipate that the size of the observed effects will be smaller than that observed by Rohrer et al. (2020). However, we do not consider the sample and geographic region to be "key ingredients" of the interleaving strategy, and we have designed our study to be sensitive to small effects that are still educationally meaningful (based on all three planned cohorts of data). Therefore, replication success on the researcher-developed measure will be determined at the end of the data collection by finding a statistically significant positive effect in favor of interleaving. Here we report on preliminary results including the first two cohorts of data collection.

Method

This study employed a classroom-based randomized design intended to isolate instructional effects while accounting for teacher-level variation. Classrooms were assigned to condition using a blocked randomization approach in which teachers served as the blocking unit. Classrooms were randomly assigned to the interleaved or blocked condition within teachers.

Teachers with a single participating class were paired with other single-class teachers from the same school or district when possible. When such pairing was not feasible, teachers were matched using district-level indicators, including geographic proximity, student–teacher ratios, and the proportion of students eligible for free or reduced-price lunch. This approach was selected to reduce imbalance in instructional context while maintaining random assignment at the classroom level. Student mathematics outcomes were assessed using a researcher-developed measure administered four weeks after the final instructional materials were delivered.

Participants

The study was conducted in sixth- and seventh-grade mathematics classrooms across multiple regions of the United States. The analytic sample represents two cohorts of an ongoing multi-cohort study. Data were drawn from 813 students (400 interleaved, 413 blocked), 51 classrooms (26 interleaved, 25 blocked) taught by 29 teachers across 15 schools across 8 school districts.

A total of 68 classes were originally recruited but 17 classes dropped (due to teacher dropout) prior to the end of the study, resulting in 25% overall cluster attrition (i.e., (68-51)/68=25%). Nine of the attrited classes were in the interleaved condition and eight were in the blocked condition, resulting in low differential attrition (2%). These levels of cluster attrition are considered low for cluster-randomized studies in education satisfying even cautious boundaries set by the What Works Clearinghouse (W.W. Clearinghouse, 2022). Demographic information for participating students is being collected from districts and was not available at the time of analysis.

Instructional Materials

Figure 1: Review Assignment design. Ten assignments (columns) contain 6 problems each (rows) of type A, B, C, or D which correspond to the targeted math skills. Items within assignments are either mostly interleaved (top) or blocked (blocked).

Review Assignments

	1	2	3	4	5	6	7	8	9	10
	Filler	A1	C1	A4	C5	D7	A8	B10	D11	A13
	Filler	A2	C2	B4	D5	C7	B8	D10	A11	Filler
Interleaved	Filler	A3	C3	D4	A6	B7	D9	C10	D12	B13
	Filler	B1	D1	C4	B6	A7	C9	A10	B12	C13
	Filler	B2	D2	B5	C6	D8	B9	C11	A12	Filler
	Filler	B3	D3	A5	D6	C8	A9	B11	C12	D13
	Filler	A1	A7	B1	B7	C1	C7	D1	D7	A13
	Filler	A2	A8	B2	B8	C2	C8	D2	D8	Filler
Blocked	Filler	A3	A9	B3	B9	C3	C9	D3	D9	B13
	Filler	A4	A10	B4	B10	C4	C10	D4	D10	C13
	Filler	A5	A11	B5	B11	C5	C11	D5	D11	Filler
	Filler	A6	A12	B6	B12	C6	C12	D6	D12	D13

Review Assignments Students completed a sequence of 10 review assignments (see Figure 1) aligned to four Common Core State Standards for sixth-grade mathematics: fraction division, decimal subtraction, rate problems, and percentage calculations, and the aligned concepts for seventh-grade mathematics. The problems were adapted from popular middle school math curricula and reviewed by math education experts to ensure their alignment to Common Core math standards. Each assignment consisted of six problems. The problems were identical across both the interleaved and blocked conditions (all students in each condition received the same problems), with the only difference being the ordering of skills which varied such that students received either interleaved practice across skills or blocked practice focusing on a single skill per assignment. To ensure the design was consistent with Rohrer et al. (2020) the problem content and sequencing structure were developed in collaboration with Dr. Rohrer.

It should be noted that by design the blocked condition of this study is not the participating school's business-as-usual context, but rather a study-specific implementation of a set of control activities which otherwise would not be implemented by participating schools. For example, under business-as-usual conditions schools might not use ASSISTments and deliver the specific problems included in the review assignments of the study. However, we chose this condition as an appropriate control because it ensures that students in both conditions had equal exposure to the math content of the study. Thus, any differences in math achievement cannot be attributed to increased practice on the target skills.

Measures

Post Assessment Student outcomes were measured using a researcher developed post-assessment - developed in collaboration with Dr. Rohrer so that it was isomorphic to the measure used in Rohrer et al. (2020) - and administered online four weeks after completion of the final review assignment, mirroring the procedure of Rohrer et al. (2020). Separate assessments were created for grade 6 and 7 to target the specific skills but used the identical development process. The assessments each consisted of 12 items designed to align with the instructional content, including three items per target skill that were aligned to the skills that were the focus of the intervention. Importantly, the assessments were not overly aligned to one of the conditions as both interleaved and blocked students reviewed the exact same items during study. Further, the assessment did not specify nor provide hints as to which strategy to apply on any problem.

Scores were z-scored within grades in order to combine both grades in analysis. Internal consistency reliability for the post-assessment was high (Cronbach's $\alpha = .84$).

Instructor Logs Logs were administered for both interleaved and blocked conditions after each review assignment. These logs contained questions about the implementation procedures to inform the fidelity and cost analysis.

Instructor Background Survey A brief background survey was administered to instructors at the beginning of the study to gather information about their teaching background and practices, in part to inform the cost analysis.

Procedure

We specified the instructional components of the intervention using a logic model (see Table 1) that links core features of an intervention with its implementation in classroom instruction (Kao, Matlen, Tiu, & Li, 2018). The logic model identifies the materials, student activities, and teacher actions associated with interleaving, review, and spacing, as well as components that were implemented similarly across conditions in this replication study. These core features were held constant to ensure consistency with the study we replicated. Pilot testing of this approach prior to the study allowed us to modify features that were not core to the intervention based on feedback. Components were defined based on the study design and established descriptions of interleaved practice. These components and their instantiation in practice were also informed by experienced teachers, administrators, and developers not participating in the study.

Implementation procedures were consistent across conditions, with the sole distinction being the ordering of content within the review assignments. Participating teachers completed a one-hour online training session tailored to the number of classes they taught. Training covered the study rationale and implementation expectations.

Table 1: Logic Model of Interleaved Practice. Core features of the intervention are indicated by an asterisk and held constant between this and the prior study which served as the basis of replication.

Intervention Component and Study Materials	Instantiation in Practice
Interleaving: Assignments created in ASSISTments that each consist of grade level appropriate problems that are interleaved	Students complete assignments that contain topics they've had prior instruction.* Students attempt each problem independently, writing out the corresponding solution steps* Students complete problems in ASSISTments in class
Review: Visual tool for showing problems and corresponding solutions, such as a whiteboard or screen	Teachers discuss problems with solutions available visually* Students instructed to correct both their work and answers* Teachers collect and verify students corrected their errors
Spacing: Ordered assignments that consist of problems of the same type spaced across the assignments	Assignments are completed in the order of the study procedure, which are distributed over time so that long-term retention benefits from the spacing effect*

Teachers were instructed to administer the pre-assessment prior to assigning any review materials, distribute the 10 review assignments sequentially with a minimum of five days between assignments, and administer the post-assessment four weeks after the final assignment. Review assignments were completed independently by students during class time. Teachers then reviewed solutions with the class, during which students revised incorrect responses on the original paper copies. Teachers were instructed to collect and review all assignments using provided solution guides and to ensure students revised incorrect solutions.

Throughout implementation, research assistants were in regular communication with teachers to support both with study tasks (e.g., remind and teachers to complete study tasks and surveys) as well as to support implementation (e.g., field implementation questions).

Cost Analysis Method

An economic evaluation of the intervention utilized the "ingredients method" for cost analysis as described in Levin et al. (2017). The ingredients method of cost estimation involves three main steps: 1) identifying and specifying the ingredients required to obtain the evaluation results, 2) gathering estimates of ingredient quantities and assigning appropriate market prices, and 3) calculating cost estimates as the product of each ingredients' quantity and price including total program costs and average cost per participant.

Table 2: Ingredients used in cost analysis and their data sources.

Ingredient	Data Source	Prices
Teacher Coordinators (research assistants that provided study support to teachers)	Interviews conducted by cost analysis experts with teacher coordinators	$52.78 per hour. per hour assuming 2,080 hours per year
Middle School Math Teachers	Regular teacher logs collected throughout implementation	75.89 per hour. Average of "Public teacher with an MA and 15-19 years of experience" and "Public teacher with a BA and 6-9 years of experience" per hour assuming 1,440 hours / year.
Chromebook	Teacher background survey collecting instructional time information, school demographic information on participating student counts by site and year.	$75.8 per hour. "Lenovo Chromebook Duet 3" total cost per unit spread evenly over 5 years of use.
Assistments Software	Study documentation of intervention	ASSISTments software is freely available

Ingredients were identified based on the logic model and discussions between cost experts and the principal investigators. These ingredients are shown in Table 2.

As noted above, the design of the control condition of this study is not the participating school's business-as-usual context. Thus, when estimating incremental costs, the difference in resource use (i.e., types and quantities of resources) between the interleaved and blocked conditions does not imply the difference between the interleaved and the participating school's business-as-usual with respect to resource use. Given this, the study team gathered information about quantities of resources allocated under business-as-usual conditions, as well as the two study conditions; interleaved and blocked. Business-as-usual data were collected primarily through the background survey administered to the second cohort of study participants. As a result, we can report on incremental costs relative to both the interleaved vs. blocked condition, and the interleaved vs. business-as-usual.

Cost estimates themselves are the product of the estimated quantities of resources—e.g., staff time and amount of Chromebook use—and each resource's market price. Table 2 summarizes the data leveraged to estimate ingredient quantities. Market prices, or the value of a resource on the open market, were assigned based on publicly available, national price data for comparable resources in educational programs. For personnel ingredients, this price includes base salary and an assigned benefit rate intended to capture benefit costs for these roles. Each salary assumes benefit rate of "State/local government employees: Educational Services. Notably, while ASSISTments is a resource allocated to both the interleaved and blocked conditions, this software is freely accessible on the open market, and thus adds no cost to implementation in either of these conditions.

Table 2 summarizes assigned prices for each ingredient. All amounts are adjusted to 2024 constant dollars and sourced from the Database of National Prices of Educational Resources published by the Cost Analysis Project, except the ASSISTments Software whose price was verified at the ASSISTments website.

Results

Impact on Math Achievement

To examine the impact of the interleaved intervention relative to blocked practice, we used a two-level hierarchical linear model to account for the nested structure of the design: students within classes, and classes blocked within teachers. Fixed effects covariates include (at the second level) dummy coded variables for the teacher blocks and grade. This model revealed that, on average, sixth and seventh grade students in the interleaved condition scored higher on a 4-week delayed post-assessment than those in the blocked condition (interleaved: M=.09, SD=1.03; blocked: M=-.08, SD=.96), which was statistically significant (Coef=.18, SE=.09, p=.047; Hedges' g=.18).

To explore whether there were differences by grade, we conducted the same model but included an interaction between grade and condition, which was not significant (p=.18).

Fidelity

Teachers reported high implementation fidelity across both conditions. All teachers administered review assignments in their prescribed order, maintaining the intended spacing structure. According to teacher logs, nearly all students in both the interleaved (92.5%) and blocked (97.2%) conditions completed assignments independently, and the vast majority attempted problems on corresponding paper worksheets (86% and 87%, respectively). All students in both conditions completed assignments during class time as intended. Teachers in both conditions recorded strong adherence to the review protocol, with most going through all solutions with students (87% interleaved, 84% blocked) and all teachers instructing students to correct both their work and answers. Teachers reported collecting and verifying student corrections at high rates in both conditions (93%). These fidelity metrics suggest that core features of the intervention—independent problem-solving, in-class completion, solution review, error correction, and spaced practice—were implemented as designed across conditions.

Cost Analysis

Cost estimates were calculated by year, site, and study condition. These raw costs were then adjusted to present value in the base implementation year, 2024, to allow more meaningful comparisons over time. Adjusted costs were then aggregated to the site level by year and study condition. Further, as is best practice, all costs are rounded to the nearest $10 or $100 to reflect the inherent imprecision in the cost estimation process. As a result of this, slight variations on cost estimates are sometimes presented as the same cost after rounding.

Overall estimated costs were very similar for the two main study conditions, with a site-level, participant-weighted average of approximately $19,200 for the Interleaved condition and $18,900 for the Blocked condition. Dividing these totals by participating students yields per-student total costs of $280 and $260 for Interleaved and Blocked conditions, respectively. Both conditions are also similar to business-as-usual conditions, which has a site-level, participant-weighted average of approximately $17,000 total and $260 per student.

The estimated costs for Interleaved and Blocked study conditions are relatively similar to each other. Furthermore, estimated costs for the Interleaved condition are not substantially different from business-as-usual condition, suggesting that implementing the interleaving strategy is unlikely to require meaningful additional investment relative to business-as-usual.

Discussion

This study reports on a partial sample of a large scale study designed to test the efficacy of interleaved practice relative to blocked practice. As the full sample is not yet collected, our conclusions are preliminary.

In the first two cohorts, we observed a statistically significant and positive effect of interleaved practice relative to blocked practice ($p < .05$, $g=.18$), thereby qualitatively replicating the positive effect of interleaved practice reported by Rohrer et al. (2020). The effect in this study is notably smaller than Rohrer et al. (.18 vs. .83). In education studies, more heterogeneous samples typically result in greater variability and thus, smaller effects (Kraft, 2020). Given that the current replication consisted of participants in multiple districts and states (relative to the single district and single state as reported in Rohrer et al), the smaller effect size is within expectation. Further, effects of 0.18 are typical for educational interventions in large studies (Lipsey et al., 2012; Kraft, 2020) and are similar to other large-scale studies on cognitive interventions (e.g., Roschelle et al. 2016; Davenport et al. 2023).

The results also suggest that the intervention can be feasibly implemented at this scale. Teacher reports indicated high implementation fidelity across both conditions, and the cost analysis suggests that implementing interleaving imposes minimal additional burden on schools. The estimated per-student cost of the interleaved condition ($280) was nearly identical to both the blocked condition ($260) and business-as-usual instruction ($260), indicating that interleaving requires little to no meaningful additional investment relative to standard practice. The low incremental cost of interleaving relative to business-as-usual is consistent with cost analyses of other technology-supported cognitive science interventions (e.g., Feng, Weiser, & Collins, 2024).

If positive effects of this report hold after the full sample of data, it would be tempting to conclude that educational practitioners and developers need only to be informed of this work to implement interleaving into their daily practice and products. This view carries a certain appeal because it suggests a straightforward fix for addressing the research practice gap. However, this view ignores the challenges of complexity and context. To scale interleaving to a broad set of contexts in the present study, we adopted a logic model and elicited the help of interdisciplinary experts. The logic model articulated core intervention features based on past research and made inferences about optimal design choices that were informed by the advice from teachers, researchers, developers. The intervention was then piloted to test its feasibility. This process is both collaborative and iterative as opposed to one shot and unidirectional. It also relies on infrastructure afforded by the current conditions of the research (e.g., grant funding) but not necessarily available to the average teacher or developer. Thus, while communicating the science of learning to educator audiences is a healthy practice for the field, more work is to be done to ensure that interleaving is implemented in the context of real world learning systems. To this end, a growing field of learning engineering dedicated to the practical application of learning research (Goodell & Kolodner, 2023) increasingly offers related approaches to support such aims (e.g., Nested learning engineering methods; Craig et al. 2024). However, there is a pressing need for additional infrastructure and incentives to support interdisciplinary collaborations between researchers, practitioners, and developers. Further, learning research could better support translation by testing educationally relevant implementation variations. The goal would be to establish design parameters that provide educators with theory-consistent guidelines for adapting learning principles to their local contexts.

Limitations

Beyond the preliminary nature of the results, several limitations warrant caution in interpreting these findings. First, as previously noted, the scope of the cost analysis does not include initial development costs. The development process involved multiple discussions, reviews, and edits to ensure the materials adhered to interleaved theory and were valid for education contexts. Further, we utilized interdisciplinary expertise to support this development process that is not included in the scope of the cost analysis. Thus, while schools or developers choosing to adopt this exact intervention could do so at low cost, schools or developers looking to interleave their own problem sets would likely face initially higher startup costs as well as practical challenges that span beyond existing infrastructure (e.g., seeking out consultations from interleaving and cognitive experts).

Second, this report includes a narrowly focused researcher-developed assessment. Although the assessment is estimated to be reliable and valid, more broadly focused tests may further attenuate effects. We plan to examine impact on students' end of year state tests but results are not available at the time of this writing. Because the existing theory predicts that we should see positive effects regardless of the assessment type, and because we have designed our study to be sensitive to small, educationally meaningful effects, we predict that we will see statistically significant positive effects on broadly focused outcome measures. However, unlike on the researcher designed measure, failure to observe statistically significant positive effects on broadly focused measures will not be considered a failure to replicate, as Rohrer et al. (2020) did not include a broadly focused measure in their study. Instead, such a finding would be interpreted as an opportunity to enhance the underlying theory - for example, if there is no effect of interleaving on this measure, it is possible that a larger dosage is required. Such hypotheses could be explored in future research.

Conclusions

At present, this study provides promising evidence that interleaving can be implemented effectively and feasibly at broad scale in education contexts. Although encouraging, age-old challenges remain with regard to translating research findings to practice. To this end, approaches that involve opportunities for interdisciplinary collaboration and iteration offer starting points for navigating contextual complexity.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305R220012 to WestEd. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

References

Carvalho, P. F., & Goldstone, R. L. (2014). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & cognition, 42(3)*, 481-495.

Cheung, A. C., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher, 45(5),* 283-292.

Craig, S. D., Avancha, K., Malhotra, P., Gorman, J. C., Verma, V., Likamwa, R., ... & Goldberg, B. (2025). Using a Nested Learning Engineering Methodology to Develop a Team Dynamic Measurement Framework for a Virtual Training Environment. In *International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale*.

Davenport, J. L., Kao, Y. S., Johannes, K. N., Hornburg, C. B., & McNeil, N. M. (2023). Improving children's understanding of mathematical equivalence: An efficacy study. *Journal of Research on Educational Effectiveness*, *16*(4), 615-642.

Kang, S. H., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. Applied Cognitive Psychology, 26(1), 97-103.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational researcher*, *49*(4), 241-253.

Feng, M., Weiser, G., & Collins, K. (2024). Cost and Cost Effectiveness of ASSISTments Online Math Support: Analysis from a Randomized Controlled Study in Middle School. *WestEd*.

Goodell, J., & Kolodner, J. (Eds.). (2022). Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond (1st ed.). Routledge. https://doi.org/10.4324/9781003276579

Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education, 24(4),* 470–497. https://doi.org/10.1007/s40593-014-0024-x

Kao, Y. S., Matlen, B. J., Tiu, M., & Li, L. (2018). Logic models as a framework for iterative user research in educational technology: illustrative cases. In *End-User Considerations in Educational Technology Design* (pp. 52-75). IGI Global Scientific Publishing.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. Educational researcher, 49(4), 241-253.

Levin, H. M., McEwan, P. J., Belfield, C., Bowden, A. B., & Shand, R. (2017). Economic evaluation in education: Cost-effectiveness and benefit-cost analysis. SAGE publications.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., ... & Busick, M. D. (2012). Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms. *National Center for Special Education Research.*

Lyle, K. B., Bego, C. R., Hopkins, R. F., Hieb, J. L., & Ralston, P. A. (2020). How the amount and spacing of retrieval practice affect the short-and long-term retention of mathematics knowledge. *Educational Psychology Review, 32(1),* 277-295.

Pomerance, L., Greenberg, J., & Walsh, K. (2016). Learning about Learning: What Every New Teacher Needs to Know. *National Council on Teacher Quality*.

Rohrer, D., & Hartwig, M. K. (2020). Unanswered questions about spaced interleaved mathematics practice. *Journal of Applied Research in Memory and Cognition*, *9*(4), 433.

Rohrer, D., Dedrick, R. F., Hartwig, M. K., & Cheung, C. N. (2020). A randomized controlled trial of interleaved mathematics practice. *Journal of Educational Psychology*, *112*(1), 40.

Rohrer, D., Dedrick, R. F., & Hartwig, M. K. (2020). The scarcity of interleaved practice in mathematics textbooks. *Educational Psychology Review*, *32*(3), 873-883.

Roschelle, J., Feng, M., Murphy, R. F., & Mason, C. A. (2016). Online mathematics homework increases student achievement. *AERA open, 2(4)*, 2332858416673968.

Stigler, J. W., Son, J. Y., Givvin, K. B., Blake, A. B., Fries, L., Shaw, S. T., & Tucker, M. C. (2020). The Better Book approach for education research and development. *Teachers College Record, 122(9)*, 1-32.

What Works Clearinghouse (2022). What Works Clearinghouse procedures and standards handbook, version 5.0. *U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance (NCEE)*. https://ies.ed.gov/ncee/wwc/Handbooks.