2023

# Determining the Quality of the Evidence Base for Incremental Rehearsal

Emily K. Fischer
*Minnesota State University, Mankato*

# Determining the Quality of the Evidence Base for Incremental Rehearsal

A Dissertation Proposal

Submitted to the Faculty of

Minnesota State University, Mankato in partial fulfillment

of the requirements

for the degree of Doctor of Psychology

by

Emily K. Fischer

Department of Psychology, Minnesota State University, Mankato

Dissertation Chair: Dr. Shawna Petersen-Brown

Date: April 2023

Title: Determining the Quality of the Evidence-Base for Incremental Rehearsal

Name: Emily K. Fischer

This dissertation has been examined and approved by the following members of the student's committee.

Shawna Petersen-Brown, Ph.D.,
Advisor

Elyse Farnsworth, Ph.D.,
Committee Member

Kyena Cornelius, Ed.D.,
Committee Member

Carlos J. Panahon, Ph.D.,
Committee Member

# Dedication

To my parents, for your continued love and support. I wouldn't be who I am today without you. To my brothers, who will doubtless be over the moon to read about flashcards, thank you for always believing in me. To my significant other, for your patience throughout this whole process. You have kept me afloat. Finally, to my grandparents, who supported me to the end. I will never forget you.

Acknowledgements

I would like to thank my committee members, Dr. Shawna Petersen-Brown, Dr. Chip Panahon, Dr. Elyse Farnsworth, and Dr. Kyena Cornelius. Your support has been invaluable throughout my graduate training. Thank you especially to my advisor, Dr. Shawna Petersen-Brown, for making this project possible, and for always pushing me to do my best. A big thank you to my cohort members, Abbey Riese, Courtney Sowle, Ariana Groen, and Collin Seifert, for letting me complain endlessly about writing a dissertation. Your encouragement and friendship have kept me going. Finally, thank you to Elizabeth Kinsey-Hawley, for volunteering your valuable time and energy to conduct inter-rater agreement for this study. This study wouldn't have been possible without you.

Table of Contents

Abstract of the Dissertation

Determining the Quality of the Evidence Base for Incremental Rehearsal

By

Emily K. Fischer, M.S.

Doctor of Psychology in School Psychology
College of Graduate Studies and Research
Minnesota State University, Mankato, 2023

The purpose of this review is to examine the current literature on incremental rehearsal (IR) to investigate whether IR is be considered an evidence-based practice, based on the quality indicators set forth by the Council for Exceptional Children (Cook et al., 2014). Burns et al. (2012) completed a meta-analysis to investigate the effectiveness and efficiency of IR and to compare the effect sizes calculated from single-case and group designs. Results of that analysis showed that IR was effective with various student groups, including students in grades ranging from preschool to high school, and students with disabilities. The original review investigated the effectiveness of IR but did not investigate the rigor of individual studies and whether IR should be considered an evidence-based practice. Given that IR is supported by a considerable body of research and has been demonstrated to be effective within that research overall, this review evaluated the existing research to determine if IR can be considered an evidence-based practice. The results indicated that IR is a practice with mixed evidence. The studies included in this review showed high methodological quality in the areas of context and setting, participants, internal validity, outcome measures, and data analysis. The studies in this review showed methodological weakness related to implementation fidelity and intervention agents.

**Determining the Quality of the Evidence Base for Incremental Rehearsal**

According to Haring and colleagues' learning hierarchy (1978), the process of learning a new skill consists of four stages. The first stage is acquisition. Acquisition is the process of learning how to complete a skill. The goal of this stage is improving the student's accuracy. The second stage is fluency. Fluency is the ability to recognize and respond to learning targets accurately and quickly. Students in the fluency stage are typically accurate but are still slow to demonstrate the skill. The goal of this stage is to improve the students' rate of response. Once students' rate of response improves, they move onto the generalization stage. The goal of the generalization stage is for students to apply what they have learned to different but relevant settings. Finally, the last stage is the adaptation stage. The goal of the adaptation stage is to adapt the newly learned skill for use in novel situations.

Although every stage is important, acquisition creates the foundation for all subsequent stages. Acquisition is specifically defined as "the period between the first appearance of the desired behavior and the reasonably accurate performance of that behavior" (Haring et al., 1978, p. 25, as cited by Daly et al., 1996). During the acquisition stage, accuracy of performance is unstable and lower than would be expected for a skill that has been mastered (Daly et al., 1996). During the acquisition stage, it is important to build accuracy through modeling, demonstration, prompting, and cueing (Haring et al., 1978, as cited by Daly et al., 1996). Acquisition precedes fluency. Without accuracy, students cannot improve their rate of response (Haring et al., 1978).

The next stage in the instructional hierarchy is fluency (Haring & Eaton, 1978). Fluency is the ability to perform a skill consistently, quickly, and accurately (Daly et al., 1996). Fluency, or rate of response, is critical because it facilitates skill building and reduces the cognitive load required to complete basic tasks. According to Laberge and Samuel's (1974) Theory of Automatic Information Processing, higher order thinking skills, such as reading comprehension, require more cognitive resources than more basic skills, such as word identification or decoding. If a student is struggling with basic skills, they will exhaust their cognitive resources and will not have enough resources left over for higher order skills. Strategies for promoting fluency include reinforcement and drilling practices (Daly et al., 1996).

One way educators can help students increase accuracy and fluency in basic skills is through flashcard interventions (Browder & Xin, 1998; Tan & Nicholson, 1997, as cited by Nist & Joseph, 2008). Specifically, traditional drill and practice methods, in which a student is shown an unknown item written on a flashcard and prompted to respond before moving onto the next unknown item, are often used. However, there are several shortcomings of traditional drill and practice methods, which include insufficient opportunities to respond to unknown items, a focus on massed practice, and a lack of behavioral momentum (i.e., a low percentage of known items; Nist & Joseph, 2008; Burns et al., 2009). An intervention called Incremental rehearsal (IR) was created to address these shortcomings. IR (Tucker, 1989) is a flashcard intervention that addresses these shortcomings using many opportunities to respond (OTRs), spaced practice, and a high percentage of known items to facilitate behavioral momentum.

## Flashcard Interventions

When using flashcard interventions, it is helpful to use methods that are more likely to be effective than traditional drill and practice. Interleaved practice and the Leitner System are two flashcard methods that, similar to IR, have features that make them more likely to be effective than traditional drill and practice. For example, these three flashcard technique share an emphasis on spaced practice.

### Interleaved Practice

Interleaving is the process of practicing different skills in an intermixed fashion rather than grouping the skills by type (Taylor & Rohrer, 2009). For example, a teacher may choose to interleave a set of math flashcards. Instead of practicing only triple digit multiplication problems, an interleaved flashcard drill would intermix single- and double-digit multiplication problems. This way, a student would have to practice all of the skills together, and would need to remember the steps to solving each type of problem, rather than remembering the steps to solving only one type of problem. An interleaved practice set would look like *abcbcacab* rather than *aaabbbccc* (Taylor & Rohrer, 2009). The beneficial effects of interleaved practice have been well studied. Interleaved practice has been found to be effective in motor skills, such as basketball shooting (Landin et al., 1993), recognition of different paintings and their artists (Kornell & Bjork, 2008), and retention of math problems (LeBlanc & Simon, 2008; Taylor & Rohrer, 2007, 2009). Interleaving practice has been shown to be more effective than "blocked" practice, in which a student practices skills grouped by type. For example, in a study on the effectiveness of interleaved practice versus blocked practice on math problems, students

who were placed into an interleaved practice group doubled their scores on a test given a day later (Taylor & Rohrer, 2009).

Interleaved practice has some drawbacks. Namely, the benefits of interleaved practice depend on the similarity between the tasks. If several tasks are easily distinguishable from one another, interleaving might be less beneficial (Taylor & Rohrer, 2009). In massed practice, tasks are easily distinguished from each other. If a student knows what they are practicing, they do not have to think about the skill needed to complete the task. For example, if a student engages in a massed practice session that involves adding unlike fractions, the student does not have to think about which formula to use. If a teacher mixed in other skills (such as adding like fractions, or multiplying fractions), the student would not only have to practice retrieving the answer, but they would also have to practice retrieving and applying the correct formula (Taylor & Rohrer, 2009). Ultimately, effectively identifying tasks that are not easily distinguishable may be difficult and less efficient for educators.

Another drawback of interleaved practice is that achieving its benefits may require a larger amount of practice. A study by Taylor and Rohrer (2009) found that massed practice facilitated perfect responding during a practice session, although interleaved practice resulted in better performance on a dependent measure on average. Additionally, the researchers found that interleaved practice required more instructional time than massed practice (Taylor & Rohrer, 2009). Finally, the success of interleaved practice requires that students possess prior knowledge of the content. It is perhaps best

for supporting fluency and generalization, since students must distinguish between several related skills to successfully use interleaved practice (Nemeth et al., 2021).

**Leitner System**

The Leitner system is another flashcard method that can be used to increase accuracy and fluency in basic skills. The Leitner System draws on the theory of spaced practice by using boxes to represent time intervals in which flashcards should be studied. In the Leitner system, students create three boxes. The first box contains all of the flashcards. These flashcards are studied regularly. Each card that is answered correctly is moved to box two. The cards in box two are reviewed every other day. Each incorrectly answered flashcard stays in box one. In review sessions including box two flashcards, these are studied before moving on to box two. During the next study session, any cards in box two that are still answered correctly get moved to box three. Flashcards in box three are studied every three to four days. During review sessions, any time a student gets a card wrong, it goes back to box one. Flashcards must progress in order from box one to box two to box three. Students can always add more boxes, but three is the minimum recommended number of boxes (Oklahoma State University, n.d.).

One benefit of the Leitner System is that it is an easy way for students to practice on their own. The boxes allow the students to clearly and efficiently monitor their own progress. Students can also change the number of days between study sessions for each box, which allows them to customize their practice schedule. One major drawback of the Leitner System is that there is not much research on it. A search of APA PsycINFO for several search terms related to the Leitner System (e.g., “Leitner System”, “Leitner Method”) did not yield any related search results.

**Incremental Rehearsal**

Finally, IR is a flashcard intervention that uses a high ratio of known items to unknown items. Typically, there are eight or nine known items for every one unknown item (Burns et al., 2012). When conducting IR, an interventionist would first present one unknown item followed by one known item. Then, the interventionist would present the unknown item and then the first two known items. After this, the interventionist would present the unknown item and the first three known items. This would continue until all known items were presented. Once all known items were presented, the last known item would be removed, and the second unknown item would be moved to the front of the stack (Petersen-Brown & Burns, 2018). Depending on the number of known items in the stack of flashcards, the student gets eight to nine opportunities to practice the unknown word. The procedure looks like this: U, K, U, KK, U, KKK, U, KKKK, U, KKKKK, U, KKKKKK, U, KKKKKKK, U, KKKKKKKK, U, KKKKKKKKK.

IR was chosen as the focus of this review rather than another flashcard method because of the focus on building accuracy and fluency, the high levels of retention associated with the use of IR, (Burns et al., 2012; MacQuarrie et al., 2002; Nist & Joseph, 2008), and the numerous research studies conducted on IR, including two meta-analyses (Burns et al., 2012; Petersen-Brown et al., 2022).

**The Role of Spaced Practice in the Effectiveness of Flashcard Interventions**

Research suggests that the spacing effect is a promising causal mechanism of all three previously described flashcard methods (Swehla et al., 2016; Taylor & Rohrer, 2009). Practicing a skill all at once, as in massed practice, has been shown to lead to better results on short-term retention tests, but practicing a skill in spaced intervals has

been found to lead to better performance on long-term retention tests (Rohrer & Taylor, 2007; Swehla et al., 2016). Interleaved practice, the Leitner System, and IR each use the spacing effect. Interleaved practice spaces intermixes different types of problems, meaning that there is a space between the same types of problems. The Leitner System spaces problems by separating the flashcards into different boxes, which are practiced on different days. IR uses the spacing effect by increasing the length of time until the unknown item is presented again. IR starts out with less spacing, resembling massed practice, and slowly adds more space between each presentation of the unknown item. The limited spacing at the beginning of the sequence results in rapid learning. The gradual increase in spacing throughout the rest of the IR sequence results in long-term retention (Swehla et al., 2016).

**Causal Mechanisms Unique to IR**

Previous research suggests that high opportunities to respond (OTRs) may be a causal mechanism of IR (Szadokierski & Burns, 2008). One study compared four different versions of IR on the acquisition and retention of the pronunciation and definition of Esperanto words. The four conditions differed on the percentage of known words (high = 90%, and moderate = 50%), and the number of OTRs (high vs. low). Using a within-subjects ANOVA, the authors found a significant main effect for OTRs, and nonsignificant effects for the ratio of known to unknown words and interactions between the variables. Additionally, increasing OTRs from low to high resulted in a large effect size ($d$ = 2.46), but increasing the ratio of known material from moderate to high resulted in a small effect size ($d$ = 0.16) (Szadokierski & Burns, 2008).

The high number of OTRs as a feature of IR contrasts with other flashcard interventions, such as drill sandwich and traditional drill and practice methods (Coulter & Coulter, 1989), in which students have fewer opportunities to respond to the unknown item. Compared to drill sandwich and traditional drill and practice, IR was found to lead to higher rates of retention in elementary and middle-school students who were tested on the acquisition and definition of Esperanto words (MacQuarrie et al., 2002). A high number of OTRs has also been shown to increase retention of sight words in students with moderate intellectual disabilities (Burns, 2007). Researchers compared IR conditions in which students were taught five sight words each day using high OTR and moderate OTR conditions, and a ratio of 10% unknown words to 90% known words. The high OTR condition led to better retention than the moderate OTR condition even though the ratio of unknown words to known words was equal (Burns, 2007). Finally, the high number of OTRs that is built into IR has been found to increase the retention of math facts in students with a specific learning disability in math computation (Burns, 2005).

Providing multiple opportunities to respond within a framework of multiple known items to few unknown items creates behavioral momentum (Burns et al., 2009). Behavioral momentum states that compliance with difficult or low probability tasks is more likely to occur if it is preceded by easier or high probability tasks. The frequency of reinforcement received by completing a series of high probability tasks makes it more likely that a student will persist when asked to complete a low probability task. Because IR involves the presentation of a high number of known items to a low number of known items, it is hypothesized that the reinforcement received from accurate completion of the

known items results in greater persistence when students are presented with the unknown items. Burns and colleagues (2009) found that students assigned to a behavioral momentum condition, in which easier words were placed at the beginning of a word reading list, read significantly more words correctly per minute than students in a control condition.

**Evidence Regarding IR's Effectiveness**

A meta-analysis that reviewed 19 studies on IR found it to be effective for various outcomes, subjects, and groups of students. Specifically, data from the single case-design studies yielded a nonoverlap of all pairs (NAP) score of 98.9% (95% CI = 97.6-100%), a large effect, with a weighted phi of .77 (95% CI = .69-.83). Data from the group design studies yielded a *d* of 1.67 (95% CI = 1.43-1.91), which converted to a weighted phi of .63 (95% CI = .39-.87), a moderate effect (Burns, 2012).

Results of the meta-analysis suggested that IR is effective across ages, populations of students, such as students in general education, and students receiving special education, type of information taught, and findings. No significant moderators were found. The 19 studies included in the review were coded according to comparison condition, stimuli used for instruction, type of assessment used, student characteristics, and efficiency of IR. Given that no moderators were significant, IR was found to be effective across levels of the potential moderators described below.

Eight studies compared IR to a baseline or control condition that received no treatment. Nine studies compared IR to a different drill condition, and two studies compared to IR to another treatment such as guided reading. As far as stimuli used for

instruction are concerned, 14 of the studies taught words with IR, two studies taught letter sounds, and two taught single-digit math facts. One study combined letter sounds and letter words. For types of assessment, nine studies used measures of recall, three involved recalling definitions or translations of words, three measured reading fluency, and two measured reading comprehension. Two studies also measured time on task.

Student characteristics were coded according to grade and disability category. 11 studies involved students in kindergarten through third grade, three studies involved students in grades four through sixth, three studies involved students in grades seven and eight, and one study included students in grades nine through twelve. Additionally, 10 studies involved students without a disability, four studies involved students diagnosed with a learning disability, three studies included students with a cognitive impairment, and one study included a student with an emotional behavioral disorder. One study involved students who were English Language Learners (ELLs). Finally, five studies examined the efficiency of IR. All except one of these studies included traditional drill and practice as a comparison condition. The efficiency of IR was implicated as an area for future research. (Burns et al., 2012).

## Evidence-Based Practices

The meta-analysis by Burns (2012) did not evaluate study rigor. IR is an intensive intervention, and extensive resources are allotted to its intervention. When extensive resources are allocated to implementation, it is important to evaluate whether or not the practice can be considered evidence-based. Evidence-based practices are defined as practices that are supported by multiple, high-quality studies that utilize research designs

from which causality can be inferred and that demonstrate meaningful effects on student outcomes (Cook & Cothren Cook, 2011).

**Elements of Evidence-Based Practices**

There are four defining elements of evidence-based practices. These elements include research design, study quality, quantity of studies, and magnitude of effect (Cook & Cothren Cook, 2011). The priority when identifying high-quality studies that support a practice is establishing strong internal validity; in other words, to establish a strong research design and study quality. Internal validity is the extent to which the observed results represent the truth and are not due to methodological errors (Patino & Ferreira, 2018). A practice cannot be considered evidence-based without strong internal validity. Internal validity is important because it allows researchers to conclude that the changes in the dependent variable were caused by changes in the independent variable. In other words, researchers can infer causality rather than correlation.

Research design is the first element in determining whether a study should be included in identifying an evidence-based practice. It is recommended that researchers only consider group experimental, group quasi-experimental, and single-subject research design studies when determining evidence-based practices (Cook & Cook, 2011). These research designs address whether the independent variable causes changes in a dependent variable, and not just whether they are correlated. These research designs are rigorous enough to rule out potential confounding variables and control for threats to internal validity (Cook & Cook, 2011).

Reviewers also assess the second element of evidence-based practices, which is study quality. Study quality means that critical features of the setting, interventionist, and participants are described. Additionally, the intervention must be well defined, and researchers must provide evidence of implementation fidelity. High quality studies are thoughtfully designed and carry out their studies so that threats to internal validity are controlled and causality can be inferred (Cook et al., 2014; Kratochwill et al., 2010).

The third element of evidence-based practices is the quantity of research studies. As multiple studies replicate findings, one can be more confident in the research findings. At least two high-quality or four-acceptable quality group experimental and quasi experimental studies must support a practice for it to be considered an evidence-based practice (Cook & Cothren Cook, 2011). In single-subject research, at least five high-quality single-subject research studies published in peer-reviewed journals, conducted in at least three different geographical locations, conducted by at least three different researchers, and including a minimum of 20 participants across studies must support the practice (Cook & Cothren Cook, 2011). Having an appropriate number of high-quality studies that control for threats to internal validity means that reviewers can infer causality from multiple sources.

Finally, the fourth element of evidence-based practices is the magnitude of effect. Evidence-based practices should demonstrate a weighted effect size that is significantly greater than zero across high and adequate quality research studies (Cook & Cothren Cook, 2011). For single-subject designs, all high-quality studies should demonstrate that the magnitude of change in student outcomes as a result of the intervention is socially

important (Cook & Cothren Cook, 2011). If the other elements of evidence-based practices are in place, reviewers can confidently infer causality. Then, the magnitude of the effect is the final consideration which establishes the social importance of the practice.

**The Importance of Implementing Evidence-Based Practices**

Under the reauthorization of the Elementary and Secondary Education Act (ESEA), the Every Student Succeeds Act (ESSA) of 2015, schools are required to teach students to high academic standards. This includes the use of EBPs. Schools receiving funds under Title I, Section 1003 (School Improvement) are required to pick EBPs that have strong, moderate, or promising evidence. All other programs receiving funds under Titles I-IV can implement EBPs with strong, moderate, or promising evidence as well as EBPs that demonstrate a rationale (ESSA, 2015).

The Individuals with Disabilities Education Act (IDEA) of 2004 states that "in implementing early intervening services, Local Education Agencies (LEAs) may carry out activities that include: professional development activities for teachers and other school staff to enable such personnel to deliver scientifically based academic instructional and behavioral interventions, including scientifically based literacy instruction, where appropriate…" It also states that LEAs may carry out activities that include "providing educational and behavioral evaluations, services and supports, including scientifically based literacy instruction. IDEA 2004 {(613(f)(2)(A)(B), via Scroggins, n.d.).

Additionally, the effect of evidence-based practices on student outcomes is clear. Providing students with evidence-based practices ensures that the practice has been rigorously reviewed. It also ensures that a practice is suitable for the needs of the child (The Iris Center, 2014). As one source put it, would you rather be given a treatment at the hospital in which multiple, well-designed studies had been completed and was found to have a large effect, or would you rather be given a treatment that only a couple of studies have been done on that has been shown to have a moderate effect (The Iris Center, 2014.). Evidence-based practices are important to use because there is already data suggesting that they improve student outcomes (Burns et al., 2017). Making sure that an intervention is supported by evidence helps to bridge the research-to-practice gap because it directly uses the existing research. It is one thing to implement an intervention that has been observed to be helpful in the past, and another thing to pick an intervention that is supported by a body of research.

**Frameworks for Evaluating the Evidence Base of Practices**

There are several methods to identify evidence-based practices, and those provided by the What Works Clearinghouse and the Council for Exceptional Children are among the most commonly used to evaluate educational practices. The What Works Clearinghouse (WWC) criteria were developed by the U.S. Department of Education's Institute of Education Services. The WWC specifies criteria that single-case and group design studies must meet to be considered high-quality and therefore contribute to the evidence base for a particular practice. The WWC uses a review process to identify all the research on a practice, assess the quality of each study, and summarize the findings.

The WWC provides reports on the quantity of studies conducted on a practice, how high-quality the studies were, and the magnitude of effects estimated by high-quality studies (WWC, n.d.). Although the WWC includes reviews on practices that have been researched with students with disabilities, most of the WWC's reviews are on practices intended to be implemented universally WWC, n.d.). Universal educational programs are developed for all students and may not be intensive or individualized enough for many students with disabilities (Cook et al., 2014).

A group of researchers developed a framework for evaluating the evidence base for educational practices on behalf of the Council for Exceptional Children (CEC; Cook et al., 2014). The CEC standards are unique in that they are made so that special education researchers can classify the evidence-base of practices on their own, and they were developed primarily for practices that are intensive and provided primarily to students with disabilities or at-risk for disabilities (CEC; Cook et al., 2014). The CEC standards evaluate factors relevant to single-case design studies and group design studies on their own, as well as factors relevant to both. The CEC standards are directly applicable to evaluating intensive interventions (Cook et al., 2014).

In terms of the four elements of evidence-based practices, each framework has different requirements. Although all frameworks mostly agree on research design, with acceptable designs including randomized controlled trials, quasi-experimental designs, ABAB designs, multiple-baseline designs, changing criterion designs, and alternating treatment designs, the WWC also includes specific standards for studies using cluster-level assignment, and the CEC does not. Regarding study quality, both frameworks focus

on elements of internal validity. Common elements that pertain to single-case design methods include at least three data points in the baseline phase that show a negative or neutral trend and at least three demonstrations of experimental effect at three different times. Common elements that pertain to group design studies include the method of assignment to groups, attrition rates, and baseline equivalence. Other aspects of study quality that both frameworks focus on that are applicable to single case design and group design studies include adequate descriptions of participants (e.g., demographics), monitoring implementation fidelity, and restricting access to treatment in control groups or baseline phases.

The third element of evidence-based practices, quantity of studies, is where the frameworks differ. The CEC states that to be considered an evidence-based practice, an intervention must be supported by at least two methodologically sound group comparison studies with random assignment to groups, four methodologically sound group comparison studies with nonrandom assignment to groups, or five methodologically sound single-subject studies (Cook et al., 2014). The WWC does not specifically state the number of studies that must be included in order to be considered an evidence-based practice. Instead, the WWC only requires one study to meet WWC standards in order to be included in the WWC database. However, the WWC reports on the extent of the evidence, or number of studies, used to determine a practice's promise. Practices with a medium to large evidence base include more than one high-quality study, and practices with a small evidence base include one high-quality study (WWC Procedures Handbook, 2017).

In terms of magnitude of effect, the CEC suggests that review teams set their own effect size criteria that must be justified based on what constitutes a socially valid level of improvement in student performance. However, they suggest using an effect size cut off of $d \geq 0.4$ = positive effects and $d \leq -0.40$ = negative effects, with neutral or mixed effects indicated by $-0.40 < d < 0.40$ for reviews of practices that target individual learners and are typically assessed using researcher-developed outcomes (Cook et al., 2014). The WWC considers effect sizes of 0.25 standard deviations or larger to be substantively important (WWC Procedures Handbook, 2017). The CEC criteria are likely to be most relevant in evaluating IR because it is a practice targeted to individual learners.

Each framework classifies practices and individual studies differently. The WWC classifies studies as meets standards, meets with reservations, or does not meet standards. The WWC also classifies practices with evidence tiers for each domain assessed in high-quality studies (WWC, 2022). After identifying the domains assessed in high-quality studies, a rating of strong, moderate, or minimal evidence is provided. The CEC classifies practices as an evidence-based practice, a potentially evidence-based practice, a practice with mixed evidence, or a practice with insufficient evidence. The CEC does not classify individual studies as methodologically sound if they meet each of the QI standards set by the CEC (Cook et al., 2014).

## Purpose and Research Questions

Implementing evidence-based practices in educational settings is important because evidence-based practices are likely to be effective when implemented with

fidelity based on empirical evidence. Individual studies may show that a practice is effective, but until a practice is verified to be evidence-based, the effectiveness of the practice is called into question as the research supporting the evidence of effectiveness may not be sound.

At this time, the body of research is generally supportive of the effectiveness of IR, but IR has not been established as an evidence-based practice because the quality of the research has not been investigated. Incremental rehearsal appears to be a useful tool for supporting skill acquisition in a variety of skill areas. Given that IR is supported by a considerable body of research and has been demonstrated to be effective within that research, it is important to evaluate the existing research to determine if IR can be considered evidence-based. Because it is a practice that shows such promise, it needs to be shown to have a high-quality or acceptable evidence base so that educators may use it confidently. The following research questions will be addressed through this proposed study:

1. Is IR an evidence-based practice according to the CEC's Standards for Classifying the Evidence Base of Practice in Special Education (Cook et al., 2014)?
2. What are the methodological strengths and weaknesses of the studies included in the review, based on the QIs set forth by the CEC (Cook et al., 2014)?

## Methods

The PsycINFO, ERIC EBSCO, ERIC ProQuest, and Academic Search Premier databases were systematically searched on May 12, 2022 to identify studies that

investigated the effectiveness of IR. Search terms included “incremental rehearsal” and “expanded” + “interspersal.” In addition, an ancestral search was conducted of other reviews of IR (Burns et al., 2012). Duplicated studies from the searches were eliminated from consideration. Manuscripts were then reviewed to determine whether they met the following inclusion criteria, adapted from Burns and colleagues’ (2012) meta-analysis:

1. The study was published in a peer-reviewed journal or was a doctoral dissertation or master’s thesis.
2. The study was an experimental group or single-case design.
3. The article was written in English.
4. The study implemented incremental rehearsal, specifically a practice method with increased spacing in between practice opportunities when an item is being taught.
5. The student receiving the intervention was school-aged (i.e., kindergarten through 12$^{th}$ grade) or receiving instruction through a school district-based program (i.e., included preschool or transition-aged students).
6. At least one of the dependent variables pertained to information learned during the intervention.

After duplicates were deleted, 99 search results remained. Next, 41 results were eliminated because they were not related to IR. Therefore, a total of 58 studies were assessed for eligibility using the inclusion criteria. Through this process, 12 studies were eliminated. Four studies were eliminated because they were studies on variations of IR, and not IR itself. Two studies were eliminated because they were literature reviews; one

study was eliminated because it was a meta-analysis; one study was eliminated because it was a "how to" guide for implementing IR, rather than a study; one study was eliminated because it was a commentary on IR, one study was eliminated because the research was not conducted in a school setting. Further, one study was eliminated because it used a version of IR as an assessment tool and not an intervention, one was about the general effects of repetition on memory. After all exclusions, 46 studies were left. The results are shown in Figure 1.

**Measures and Materials**

A rubric was created and implemented based on the CEC's Standards for Classifying the Evidence-Base of Practices (Standards; Cook et al., 2014). The rubric was adapted from rubrics used in the application of the CEC's Standards in previous investigations (Jitendra et al., 2011; Petersen-Brown et. al., 2021). The rubric for this investigation was modified to reflect the current Standards, and to pertain to IR. The rubric included the following eight overall quality indicator (QI) categories and 28 QIs within those categories. Most QIs applied to both single-case design and group design studies. However, some QIs only applied to one or the other. See Appendices A-D for the rubric and coding forms.

- **Context and setting:** This category included one QI and evaluated details provided regarding where the research was conducted. The QI was applied to single-case and group design studies.
- **Participants:** This category included two QIs and evaluated information provided regarding the participants, in particular the criteria for selecting

participants (i.e., identification of a disability or skill deficit). Both QIs were applied to single-case and group design studies.

- **Intervention Agent:** This category included two QIs and reviewed details related to the role of the intervention agent and specific training or qualifications held by the intervention agent. Both QIs were applied to single-case and group design studies.
- **Description of Practice:** This category included two QIs and evaluated information related to the intervention procedures (i.e., intervention components, instructional behaviors, dosage) and if applicable, materials needed to implement the intervention. Both QIs were applied to single-case and group design studies.
- **Implementation Fidelity:** This category included three QIs and evaluated information related to adherence using direct, reliable measures, fidelity related to dosage or exposure, and the degree to which the study assessed and reported implementation fidelity regularly throughout implementation and for each interventionist, setting, participant, or other unit of analysis. The QIs were applied to single-case and group design studies.
- **Internal Validity:** This category included nine QIs. The QIs evaluated information related to control and systematic manipulation of the independent variables, descriptions of baseline or control conditions, assignment to groups, demonstrations of experimental effect, control over common threats to internal validity, and overall and differential attrition

rates. Three QIs were applied to single-case and group design studies, three QIs were applied to group design studies, and three QIs were applied to single-case design studies.

- **Outcome Measures:** This category included six QIs. The QIs evaluated information related to the social importance of outcomes, a clear definition and measurement of the dependent variables, the effects of the intervention on all measures of the outcome targeted by the review, the frequency and timing of the outcome measures, adequate evidence of reliability, and adequate evidence of validity. Five of the QIs were applied to single-case and group design studies and one QI was applied to group design studies.
- **Data analysis**: This category included three QIs. The QIs evaluated information related to appropriate data analysis techniques, one or more appropriate effect size statistics, and in the case of single-case design studies, a single-subject graph that clearly represents outcome data across all study phases for each unit of analysis. Two QIs were applied to group design studies and one QI was applied to single-case design studies.

Each QI was rated using a 0, indicating a QI was not met, a1, indicating a QI was met, or n/a, indicating a QI was not applicable In the case of the QI "Evidence of validity is provided and sufficient" under the Outcome Measures QI category, n/a was an option when the dependent variable for a study was retention or efficiency, because these constructs are measured using researcher-derived methods, and therefore evidence of

validity cannot be provided. In the case of "Baseline phase has three datapoints" under the Internal Validity QI, n/a was an option when an alternating treatment design without a baseline phase was implemented.

A mean rating was calculated for each QI category for each study. The rubric criteria were expanded and operationalized to facilitate agreement and clarity in how specific QIs were met or not met. Twenty four of the 28 subcategories of QIs were applied to group design studies, and 22 were applied to single-case design studies. Rubric ratings were entered onto a spreadsheet to facilitate efficient data management.

## Procedures

Procedures included the training process, the rating process, and the process of determining the quality of the evidence base for IR.

### *Training and Rating Process*

An advanced graduate student in school psychology, who was familiar with IR, provided interrater agreement (IRA) on inclusion, rating, and effect size. First, the graduate student was trained in the inclusion criteria by the author during one 30-min training session. After the training session, the graduate student independently reached 90% IRA on two studies. At this point, the graduate student continued to code studies for inclusion. A random number generator was used to identify 20 search results of the 58 to evaluate for inclusion, and the graduate student rater evaluated these for inclusion. Agreement for inclusion was 100%.

Once studies were coded for inclusion, and the rating process began, the author trained the graduate student in the rating process. A 60-min training session was held in

which the rubric and the coding forms were introduced. The author modeled the rating process for one single-case design study and one group design study. Then, the graduate student rater rated a different single-case design study and a different group design study with assistance and feedback provided by the author. Following this, the rater rated one group design study and one single-case design study independently. Feedback was provided at a second session, and agreement was calculated prior to discussing and resolving disagreements. At the second session, agreement was 90% before disagreements were resolved. Two additional studies, one group design study and one single-case design study, were then provided to the rater, and they rated the studies independently. At a third session, agreement was calculated prior to discussing and resolving disagreements. At this session, agreement was 100%. The goal for agreement was 90% agreement with the author for two studies in a row for both group and single-case design studies. As the goal was met, the rater began rating studies independently. After the training process was complete, the author rated the included studies using the group design and single-case design rubric and coding forms. Throughout the coding process, 16 studies (33%) were randomly selected and coded by the author and the graduate student rater to facilitate IRA. Seven (43%) single-case design studies were included, and nine (56%) group design studies were included. At periodic meetings, the author and the rater met to discuss and resolve any disagreements. Re-training occurred one time, when average agreement fell below 90%. Average IRA was 93%, with a range of 79-100% for each study.

### *Data Analysis*

The magnitude of effect was calculated for studies that were classified as methodologically sound and that compared IR to an intervention that was different from IR. If a study compared IR to another version of IR, an effect size was not calculated. Studies were classified as having positive, neutral/mixed, or negative effects. For group comparison studies, the effect size was based on the hinge point of $d = 0.4$ (Hattie, 2009). When $d$ was greater than or equal to 0.4, the study was considered to have positive effects. When $d$ was between –0.4 and 0.4, the study was considered to have neutral or mixed effects. When $d$ was equal to or less than –0.4, the study was considered to have negative effects. Effect sizes were calculated using the version of IR hypothesized to be most effective as compared to the non-IR condition hypothesized to be the least effective (i.e., the condition containing the fewest active elements). Cohen's d was used as an estimate of effect size (Cohen, 1988).

Single-case design studies were considered to have positive, neutral or mixed, or negative effects on the basis of the number and proportion of participants in a study for whom a functional relationship between the independent variable and dependent variable was established, and the direction of the functional relationship (Cook et al., 2014). Functional relationships were established by visual analysis, including analysis of changes in level, trend, variability, immediacy of effect, overlap of data points across phases, and replication of effects (Cook et al., 2014). Evidence for replication of effects was not mentioned by Cook et al. (2014) as being necessary to determine the effect of a single-case design study, however, Cook et al. (2014), cites previous CEC standards in

which evidence of replication was necessary to determine a functional relationship (Horner et al., 2005).

Single-case design studies were considered to have positive effects when a functional relationship was established between the independent and dependent variables, resulting in a meaningful change in the dependent variable for at least 75% of the cases in a study. Additionally, to be considered as having a positive effect, there had to be a minimum of three total cases, and the data for none of the cases showed evidence of a negative functional relationship between the independent variable and the dependent variable (Cook et al., 2014). A single-case design study was considered to have negative effects when a functional relationship was established between the independent and dependent variables resulting in a nontherapeutic change in the targeted dependent variables for at least 75% of relevant cases in a study. Again, there had to be a minimum of three cases (Cook et al., 2014). A single-case design study was considered to have neutral or mixed effects when the criteria for neither positive or negative effects were met.

Similar to the process used for conducting IRA on QIs, IRA was conducted on one (50%) methodologically sound single-case design study and one (50%) group design study from which an effect size could be calculated. A graduate student rater met with the author in a 30-min training session on determining effect size. The graduate student rater independently determined the effect size for one single case design study and one group design study. 100% IRA was met, and so the graduate student independently determined the effect size for the other studies. Overall IRA for effect size was 100%.

### *Determining Evidence-Based Practice*

After the methodological soundness of each study was determined based on the rubrics and coding forms, the evidence base of IR as a practice was determined. Based on the Standards (Cook et al.; 2014), a practice was considered evidence-based if it was supported by:

(a)

- Two methodologically sound group comparison studies with random assignment to groups, positive effects, and at least 60 total participants across studies;
- Four methodologically sound group comparison studies with nonrandom assignment to groups, positive effects, and at least 120 total participants across studies; or
- Five methodologically sound single-subject studies with positive effects and at least 20 total participants across studies; or

(b) Met at least 50% criteria for two or more of the study designs described in (a). For example, the practice was supported by:

- One methodologically sound group comparison study with random assignment, positive effects, and at least 30 total participants, as well as three methodologically sound single-subject research studies with positive effects and at least 10 total participants; or
- Three methodologically sound single-subject studies with positive effects and at least 10 total participants, as well as two methodologically sound

group comparisons studies with nonrandom assignment, positive effects, and at least 60 total participants; AND

(c)

- Included no methodologically sound studies conducted with negative effects and at least a 3:1 ratio of methodologically sound studies with positive effects to methodologically sound studies with neutral/mixed effects.

A practice was considered potentially evidence-based if it was supported by:

(a)

- One methodologically sound group comparison study with random assignment to groups and positive effects.
- Two or three methodologically sound group comparison studies with nonrandom assignment to groups and positive effects; OR

(b) Met at least 50% of criteria for two or more of the study designs described in (a) AND

(c) Included no methodologically sound studies conducted with negative effects, and at least a 2:1 ratio of methodologically sound studies with positive effects to methodologically sound studies with neutral/mixed effects.

IR was considered to have mixed evidence if:

(a) Criterion (a) or (b) for evidence-based practice or potentially evidence-based practice (regarding the number of methodologically sound studies with positive effects supporting the practice) was met, AND

(b) The ratio of methodologically sound studies with positive effects to methodologically sound studies with neutral/mixed effects was less than 2:1; OR one or more methodologically sound studies was conducted with negative effects, as long as methodologically sound studies with negative effects did not outnumber methodologically sound studies with positive effects.

IR was considered to have insufficient evidence if insufficient research exists to meet the criteria for any of the evidence-based categories. Finally, IR was considered to have negative effects if:

(a) More than one methodologically sound study (of any acceptable design) was found to have negative effects AND

(b) The number of methodologically sound studies conducted with negative effects outnumbered the number of methodologically sound studies with positive effects.

## Results

The QI ratings of the 46 studies included in this review were compiled and are displayed in Tables 1 and 2. These ratings were used to identify five methodologically sound studies to answer the first research question. The ratings of all 46 studies were summarized to answer the second research question.

### Is IR an Evidence-Based Practice?

The first research question asked if IR is an evidence-based practice according to the CEC's Standards (Cook et al., 2014). Ultimately, IR was found to be a practice with mixed evidence. To determine this, the methodological soundness of the studies included in the review was evaluated. Five group design studies and two SCD studies were found

to be methodologically sound because they met all applicable QIs. The studies that didn't meet all the applicable QIs were not included. The group design studies included Burns et al. (2019), Joseph and Schisler (2007), Petersen-Brown and Burns (2011), Petersen-Brown and Burns (2018), and Zaslofsky et al. (2016). Two SCD studies were found to be methodologically sound: Burns (2005) and Volpe et al. (2021).

### *Analysis of Group Design Studies*

To classify the evidence base according to group design research, effect sizes were calculated to estimate the magnitude of IR's effect. Of the five studies, effect sizes were calculated for the Burns et al. (2019) study and the Joseph and Schisler (2007) study to determine the favorability of the evidence. These studies included 89 studenst total. Effect sizes were not calculated for the other three studies because they did not compare IR to a non-IR comparison condition (rather, they compared to a variation of IR).

Effect sizes were computed for the other studies using the version of IR hypothesized to be the most effective as compared to the non-IR condition that was designated as the control condition or was hypothesized to be the least effective. The Burns et al. (2019) study assessed retention of multiplication facts for students in an IR condition compared to students in a traditional drill and practice condition. There were 29 total participants. The effect size was positive ($d$ = 1.55), with IR having a large effect on retention of multiplication facts.

The Joseph and Schisler (2007) study compared the instructional effectiveness of students in an IR condition versus students in a traditional drill and practice condition. Instructional effectiveness was measured by each group's mean cumulative oral reading

fluency scores. There were 60 total participants. IR had a neutral/mixed effect size as compared to traditional drill and practice ($d$ = -0.05).

The Petersen-Brown and Burns (2011) study compared the effects of IR and IR with a vocabulary component on the retention and generalization of unknown words with a total of 61 participants. The Petersen-Brown and Burns (2018) study compared the effects of IR, IR with a vocabulary component, and IR with contextual reading on the maintenance and generalization of unknown word with a total of 41 participants. The Zaslofsky et al. (2016) study compared the effects of four versions of IR that varied opportunities to respond and generation effects on retention of multiplication facts. There were 104 total participants.

***Analysis of SCD Research***

Two single-case design studies were determined to be methodologically sound that included seven total participants. The first study by Burns (2005) used a multiple baseline design to look at the effects of IR on the fluency of single-digit multiplication facts in three children with learning disabilities in math computation. The effects on fluency, as measured by digits correct per min, were positive for all participants based on visual analysis. The second study by Volpe et al. (2011) used an alternating treatments design to compare the effects of IR to traditional drill and practice on the retention of unknown words for four participants. For this study, it was determined that IR did not have a meaningful effect on retention. When opportunities to respond were held constant, three out of the four students in the Volpe et al. (2011) study retained more words in the traditional drill and practice condition than in the IR condition, with a significant degree

of overlap observed. When time was held constant, three out of four students in the IR condition read more words correctly in a next day retention probe than students in the traditional drill and practice condition. However, differences across conditions were small for two of the participants (Volpe et al., 2011). Therefore, the effect of IR based on this study was determined to be neutral/mixed.

### *Level of Evidence*

Using the guidelines provided in the Standards (Cook et al., 2014), IR was first determined to be potentially evidence-based. As described above, there were seven total studies that met all the QIs and were determined to be methodologically sound according to the guidance provided in the Standards (Cook et al., 2014). Of the five methodologically sound group design studies, two permitted the computation of effect sizes comparing IR to a control or comparison condition. One included random assignment and found neutral effects (Joseph & Schisler, 2007) and one included nonrandom assignment and found positive effects (Burns, 2019). Of the two methodologically sound SCD studies, one demonstrated positive effects across each of three cases (Burns, 2005), and one demonstrated no functional relationship between IR and improved outcomes in three of four (75%) cases. Therefore, the ratio of studies identifying positive effects to studies identifying neutral/mixed effects was 1:1. Based on the description outlined previously in the Method, IR is considered a practice with mixed evidence.

**Methodological Strengths and Weaknesses of Studies Included in the Review**

To answer the second research question, the overall means for each QI and QI category were compiled. The means for group design studies and single-case design studies were also calculated independently. This information is displayed in Table 3. Overall, the studies were strongest in the Description of Practice QI category; the overall mean was 1. The second strongest QI category was the Context and Setting QI category with an overall mean of 0.98. The mean for the group design studies was 1, while the mean for the single-case design studies was 0.96. The next area of methodological strength was the Data Analysis QI category. The overall mean for this QI category was 0.93. The mean for the group design studies was 1, and the mean for the single-case design studies was 0.89. Overall, the studies in this sample provided comprehensive descriptions of the IR procedure and necessary materials and important attributes of the context of the research. Additionally, data analysis procedures were generally considered appropriate.

An area of methodological weakness was the Intervention Agent QI category. The overall mean for this QI category was 0.65 with the group design mean of 0.75 and the single-case design mean of 0.59. The first QI in this category was related to a description of the intervention agent's role. The overall mean for this QI was 0.65. The single-case design mean was 0.61, and the group design mean was 0.72. The second QI in this category was related to a description of the training and qualifications obtained by the intervention agent. The overall mean for this QI was 0.65. The single-case design mean was 0.57, and the group design mean was 0.78. This indicates that studies, especially

single case design studies, have not documented important attributes of intervention agents, including whether they have been adequately trained to implement IR.

Finally, the area that displayed the most methodological weakness was the Implementation Fidelity QI category. The overall mean was 0.61. The group design mean for this QI category was 0.57. The single-case design mean was 0.63. Overall, the third QI in this category, "as appropriate, the study assess and reports implementation fidelity (a) regularly throughout the intervention, and (b) for each interventionist, each setting, and each participant or other unit of analysis" had the lowest mean of 0.41. The single-case design studies had a mean of 0.39 for this QI. The group design studies had a mean of 0.44 for this QI. The second QI in this category, "the study assesses and reports implementation fidelity related to dosage or exposure using direct, reliable measures" had an overall mean of 0.54. The group design studies had a mean of 0.50. The single-case design studies had a mean of 0.57. The first QI in this category, "The study assesses and reports implementation fidelity related to adherence using direct, reliable measures" was the most methodologically strong of the three QIs in the category. The overall mean was 0.87. The mean for the group design studies was 0.78. The mean for the single-case design studies was 0.93. Therefore, while most studies in the sample directly assessed and reported implementation fidelity related to adherence, the fewer did so related to dosage. The minority of studies examined fidelity related to adherence and dosage across units of analysis.

There were three QI categories that were methodologically strong overall, yet the means for the group design studies and the mean for the single-case design studies

differed noticeably for certain QIs within the category. The first category was the Participants category. The overall mean for the participants category was 0.91. The mean for the group design studies was 0.83 and the mean for the single case design studies was 0.96. The QI with the largest difference between group design and single-case design studies was the second QI, “disability or risk status described.” The QI for the group design studies was 0.72, and the QI for the single-case design studies was 1. The first QI “demographics described” had an overall mean of 0.93. The group design mean was 0.94, and the single-case design mean was 0.93. Therefore, group design studies generally provided less information on the risk status of the participant sample.

The second category in which the means for the group design studies and the single-case design studies differed was the Internal Validity category. The overall mean for the Internal Validity category was 0.91. The mean for the group design studies was 0.99. The mean for the single-case design studies was 0.86. The lowest mean for the single-case design studies was 0.61, under the QI “baseline phase has at least three data points.” This QI did not apply to group design studies. The second lowest mean for the single-case design studies was 0.71, which was applied to the QI “Design controls for threats to internal validity.” This QI did not apply to group design studies. Finally, the next lowest QI for the single case design studies has a mean of 0.79 and was applied to the QI “three demonstrations of effect at three different times.” This QI did not apply to group design studies. This suggests that single-case studies were less likely to have strong internal validity, which calls into question conclusions that can be drawn from that sample of studies overall.

Finally, the Outcome Measures/Dependent Variables QI category was methodologically strong overall. The overall mean was 0.93. The mean for the group design studies was 0.91 and the mean for the single-case design studies was 0.94. However, it should be noted that one QI within this category showed distinct evidence of methodological weakness. The QI was "evidence of validity is provided and sufficient" and only applied to group design QIs. The mean for that QI was 0.38. It was only applied to group design studies that included measures beyond learning and retention of taught items. This QI was considered "not applicable" for studies that only assessed taught content, as evidence of validity is difficult to obtain on researcher-created materials for taught items. Many studies did not provide evidence that the measures they used to assess constructs (such as generalization of information learned through IR) were valid.

## Discussion

Based on the CEC's Standards (Cook et al., 2014), IR was found to be a potentially evidence-based practice with mixed evidence. This review included 46 studies, 28 were single-case design studies and 18 were group design studies. Of the included studies, seven studies were found to be methodologically sound. Of these seven studies, five were group design studies and two were single-case design studies. One of these studies found positive effects of IR, and the other found mixed/neutral effects. Effect sizes could only be calculated for two of the five group design studies found to be methodologically sound. One of these studies found positive effects of IR, and the other found mixed/neutral effects. These findings led to the determination that IR is a practice with mixed evidence.

The second research question pertained to the methodological strengths and weaknesses of the research. The sample of studies included sufficient descriptions of IR methods and the materials used (i.e., Description of Practice). These results suggest that researchers and/or educators reviewing these studies would be able to replicate these procedures in research and/or practice. Nearly all studies included sufficient information regarding the setting in which the research was conducted (i.e., Context and Setting). This suggests that researchers and/or educators reviewing these studies would be able to determine if IR has been attempted in a setting similar to their own. Nearly all studies included sufficient information regarding participant demographics (i.e., Participants). This means that educators and/or researchers would be able to determine whether IR has been found to be effective with students that share important characteristics to their own students and/or participant sample. In almost all studies, researchers manipulated the independent variable, sufficient information regarding the control/comparison conditions was described, and the control/comparison participants did not have access to treatment (i.e., Internal Validity). However, the means for some of the QIs for the single-case design studies in the QI category were low. Particularly in the QI stating "the baseline phase has three different data points and establish a pattern that predicts undesirable performance." This suggests that researchers and/or educators reviewing these studies could be confident that confounds were minimal for group design studies, but not necessarily single-case design studies. Researchers and/or educators should know that single-case design studies that they review may face threats to internal validity. Almost all studies included information on the social importance of the outcomes, an adequate

description of the dependent variables, sufficient information on the effects of all relevant measures, the frequency and timing of measurement, and evidence of reliability (i.e., Outcome Measures). These results suggest that the outcome measures used were generally of acceptable quality and appropriately measured the effects of IR. Finally, almost all studies used data analysis techniques appropriate for the study design. These results suggest that researchers and/or educators reviewing these studies could use the data to understand the magnitude of the effect of IR compared to other flashcard interventions.

Most of the studies did not include sufficient information on implementation fidelity, particularly related to dosage or exposure. Related to dosage, in most cases consumers of the studies cannot conclude that participants received the intended dosage of IR. Additionally, most of the studies did not include sufficient information on the role of the intervention agent and the intervention agent's qualifications. This may lead the reader to question the qualifications and/or training of interventionists in the research as well as the required qualifications and/or training required to implement IR. This study highlights the importance of following guidelines for ensuring methodological quality of research. Of the 46 studies included in this review, seven met the criteria set forth by the CEC, and four were considered in deciding the favorability of the evidence. Ultimately, IR was determined to be a potentially evidence-based practice with mixed evidence, even though many studies on the practice exist. These findings are concerning because the benefits of IR on building fluency and accuracy in basic skills have been widely disseminated, yet the evidence base is not strong enough to suggest the implementation of

IR without caution. Perhaps the regular, systematic use of best practice guidelines when conducting research should be more widely implemented. Journal editors should use best practice guidelines when making decisions about publication.

**Limitations**

The results of this review should be considered within the context of its limitations. The search procedures used may have inadvertently excluded some research. Some research was purposely excluded to reduce the potential for duplication of studies and to ensure the sample of studies had been subjected to a review process, either peer review or committee review.

This study involved the application of a specific framework for evaluating methodological quality, of which there are several. Utilizing an alternate framework may have yielded a different result. However, similar reviews in the area of academic intervention research have applied the CEC's Standards (Petersen-Brown et al., 2021; Cook et al., 2020; Jitendra et al., 2015). Based on the nature of IR as an intensive intervention meant for small groups of students or individual students, it was determined that the CEC framework was the most appropriate choice. Next, the application of QIs is a subjective process. However, the rubric used for this review was an adaptation of a rubric used in a prior review (Petersen-Brown et al., 2021). In addition, the IRA in this review was favorable, indicating that the rubric was applied objectively to investigate the evidence-base of IR using the CEC's Standards.

**Implications for Research**

This study suggested where researchers investigating IR should focus their efforts to ensure they are conducting methodologically sound research. The QI category of Implementation Fidelity had the lowest mean for both group and single-case design studies. In particular, QIs related to measuring and reporting fidelity related to dosage were particularly low. Based on this information, researchers should measure and report implementation fidelity related to both dosage and adherence throughout their studies. For the most part, implementation fidelity was assessed through adherence to an intervention protocol. However, study authors rarely included explicit information on the actual dosage of IR, such as in weeks or months, sessions per week, or minutes per session. Study authors also rarely included information on dosage or adherence across cases, conditions, and/or groups in a study.

The QI category of Intervention Agent also had a low overall mean. Researchers should collect and include information on key attributes of the intervention agent in research studies. The description should include information about the intervention agent's role, such as a researcher, teacher, paraprofessional, etc., and as relevant, any background variables such as educational background or licensure. Additionally, any training that the intervention agent undertook or any qualifications the interventionist received to implement the intervention should be clearly documented.

Future research may also review the seven methodologically sound studies in greater detail. A previous meta-analysis did not identify any moderators (Burns et al., 2012), finding that IR is effective in many situations. However, given that this study found

mixed evidence in a sample of methodologically sound studies, an in-depth analysis of these studies may provide helpful insight into the specific situations when IR is most likely to be effective.

**Implications for Practice**

This study suggests that educators should use IR cautiously, as it is a potentially evidence-based practice with mixed evidence. Methodologically sound studies on IR found a mix of positive and neutral results, which indicates that educators using IR should carefully monitor outcomes. Educators should keep in mind that many studies done on IR did not report fidelity related to dosage, so the effect of dosage on outcomes is unclear. Additionally, educators should remember that the role of the intervention agent, and any training or qualifications that person received were also rarely reported, so the needed qualifications and training needed to implement IR is unclear. Educators should bear this in mind as they make decisions about who the interventionist conducting IR will be. Educators should consider starting with other interventions that target acquisition and fluency of basic skills that are evidence-based in order to increase the likelihood of positive student outcomes.

**Conclusions**

IR is a practice with mixed evidence. The studies investigating IR had many methodological strengths as well as several areas for improvement. The framework that was applied to this review is rigorous, but the quality indicators are each important for ensuring the internal and external validity of research, in addition to the extent to which researchers and educators can confidently apply the findings. Researchers should

consider these frameworks when planning research, and consumers of research should interpret findings within a lens that considers the methodological quality of the study. Intervention research in the schools is unpredictable, but the compromises and adaptations that are often necessary must be balanced with maintaining methodological quality to meaningfully inform future practice. The current study identified 46 studies on IR, dating back to 1999 (Burns,1999). Given the quantity of research on IR, many researchers and practitioners may be under the impression that IR is evidence-based. It is staggering then, to realize that out of 46 studies, only seven were considered methodologically sound. This finding underscores the importance of the quality of research done on a practice over the quantity. In the end, it is not so much the amount of research done on a practice, but the caliber of the research done on a practice that matters.

## References

Adams, S. R., & Maki, K. E. (2020). Examining the differential effectiveness and efficiency of alternative multiplication drill interventions with third-grade students. *Journal of Applied School Psychology,* 1-25. https://doi.org/10.1080/15377903.2020.

Browder, D. M., & Xin, Y. P. (1998). A meta-analysis and review of sight word research and its implications for teaching functional reading to individuals with moderate to severe disabilities. *The Journal of Special Education, 29,* 400-413.

Burns, M.K. (1999). Test-retest reliability of individual student acquisition and retention rates as measured by instructional assessment. *Dissertations.* 255.

Burns, M.K. (2005). Using incremental rehearsal to increase fluency of single-digit multiplication facts with children identified as learning disabled in mathematics computation. *Education and Treatment of Children, 28*(3), 237-249.

Burns, M.K. (2007). Comparison of opportunities to respond within a drill model when rehearsing sight words with a child with mental retardation. *School Psychology Quarterly, 22*(2), 250-263.

Burns, M.K., Ardoin, S.P., Parker, D.C., Hodgson, J., Klingbeil, D.A., & Scholin, S.E. (2009). Interspersal technique and behavioral momentum for reading word lists. *School Psychology Review, 38,* 428-434.

Burns, M.K., Riley-Tillman, T.C., & Rathvon, N. (2017). *Effective School Interventions: Evidence-Based Strategies for Improving Student Outcomes* (3rd ed.). New York, NY: Guilford Press.

Burns, M.K., Zaslofsky, A.F., Kanive, R., & Parker, D.C. (2012). Meta-analysis of incremental rehearsal using phi coefficients to compare single-case and group designs. *Journal of Behavioral Education, 21,* 185-202. DOI 10.1007/s10864-012-9160-2

Chen, O., Paas, F., & Sweller, J. (2021). Spacing and interleaving effects require distinct theoretical bases: A systematic review testing the cognitive load and discriminative-contrast hypotheses. *Educational Psychology Review, 33,* 1,499-1,522.

Cohen J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Routledge Academic [Google Scholar]

Cook, B.G., Buysse, V., Klinger, J., Landrum, T.J., McWilliam, R.A., Tankersley, M., & Test, D.W. (2014). CEC's standards for classifying the evidence base of practices in special education. *Remedial and Special Education, 36*(4), 1-15. DOI:10.1177/0741932514557271

Cook, S. C., Collins, L. W., Morin, L. L., & Riccomini, P. J. (2020). Schema-based instruction for mathematical word problem solving: An evidence-based review for students with learning disabilities. *Learning Disability Quarterly, 43*(2), 75-87. https://doi.org/10.1177/0731948718823080

Cook, B.G., & Cook, S.C. (2011). Unraveling evidence-based practices in special education. *The Journal of Special Education,* 1-12. DOI: 10.1177/0022466911420877

Coulter, W. A., & Coulter, E. M. (1989). *Curriculum-based assessment for instructional design: Trainer's manual.* (Unpublished training manual available from Directions and Resources, P.O. Box 57113, New Orleans, LA 70157)

Every Student Succeeds Act, 20 U.S.C. § 6301 (2015). https://www.congress.gov/bill/114th-congress/senate-bill/1177

Daly, E.J. III, Lentz, F.E., Jr., & Boyer, J. (1996) The instructional hierarchy: A conceptual model for understanding the effective components of reading interventions. *School psychology Quarterly, 11*(4), 369-386. https://doi-org.ezproxy.mnsu.edu/10.1037/h0088941

Dempster, F. N. (1991). Synthesis of research on reviews and tests. *Educational Leadership, 48,*71-76

Haring, N.G., Lovitt, T.C., Eaton, M.D., & Hansen, C.L. (1978). *The fourth R: Research in the classroom.* Columbus, OH: Merrill

Hattie, J. (2009). Visible learning: *A synthesis of over 800 meta-analyses relating to achievement.* New York, NY: Routledge.

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

Individuals with Disabilities Education Act, 20 U.S.C. § 1400 (2004). https://sites.ed.gov/idea/

Jitendra, A., Burgess, C., & Gajria, M. (2011). Cognitive strategy instruction for improving expository text comprehension of students with learning disabilities:

The quality of the evidence. Exceptional Children, 77(2), 135–159. https://doi.org/10.1177/001440291107700201

Jitendra, A. K., Petersen-Brown, S., Lein, A. E., Zaslofsky, A. F., Kunkel, A. K., Jung, P., & Egan, A. M. (2015). Teaching mathematical word problem solving: The quality of evidence for strategy instruction priming the problem structure. *Journal of Learning Disabilities, 48*(1), 51-72. https://doi.org/10.1177/0022219413487408

Kratochwill, T.R., Hitchcock, J., Horner, R.H., Levin, J.R., Odom, S.L., Rindskopf, D.M., & Shadish, W.R. (2010). Single-case design technical documentation. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "Enemy of Induction"? *Psychological Science*, 19, 585– 592.

LaBerge, D., & Samuels, S.J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6*(2), 293-323. https://doi.org/10.1016/0010-0285(74)90015-2

Landin, D. K., Hebert, E. P., & Fairweather, M. (1993). The effects of variable practice on the performance of a basketball skill. *Research Quarterly for Exercise and Sport*, 64, 232– 236.

Le Blanc, K., & Simon, D. (2008). Mixed practice enhances retention and JOL accuracy for mathematical skills. *Paper presented at the 49th Annual Meeting of the Psychonomic Society, Chicago, IL. November, 2008*

Leitner, S. (1972). *So lernt man lernen.* Herder.

MacQuarrie, L.L., Tucker, J.A., Burns, M.K., & Hartman, B. (2002). Comparison of retention rates using traditional, drill sandwich, and incremental rehearsal flash card methods. *School Psychology Review, 31*(4), 584-595.

Nemeth, L., Werker, K., Arend, J., & Lipowsky, F. (2021). Fostering the acquisition of subtraction strategies with interleaved practice: An intervention study with German third graders. *Learning and Instruction, 71,* 11. https://doi.org/10.1016/j.learninstruc.2020.101354

Nist, L., & Joseph, L.M. (2008). Effectiveness and efficiency of flashcard drill instructional methods on urban first-graders' word recognition, acquisition, maintenance, and generalization. *School Psychology Review, 37*(3), 294-308.

Oklahoma State University (n.d.). Memory Techniques: Leitner Method. https://universitycollege.okstate.edu/lasso/site_files/documents/leitner_method.pdf

Patino, C.M., & Ferreira, J.C. (2018). Inclusion and exclusion criteria in research studies: Definitions and why they matter. *Jornal Brasileiro de Pneumologia, 44*(2). doi: 10.1590/S1806-37562018000000088

Petersen-Brown, S., Johnson, M.E., Bowen, J., Lundberg, A.R., Nelson, J.D., & Wiswell, J.M. (2021). Is repeated reading evidence-based? A review of the literature. *Preventing School Failure, 65*(4)*,* 379-391. DOI: 10.1080/1045988X.2021.1934376

Petersen-Brown, S., Kinsey Hawley, E., Fischer, E.K., Dela Paz, I.N., German, D. (2022). *The effectiveness of incremental rehearsal and implications for schools.*

[Presentation given at the National Association of School Psychologists Annual Conference, PowerPoint Slides].

Pruzan, T. (2008). *The clumsiest people in Europe.* Bloomsbury Publishing.

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics practice problems boosts learning. *Instructional Science*, 35, 481– 498.

Scroggins, L. (n.d.). *Evidence-based practices in education.* Office of the State Superintendent of Education. Washington, D.C. https://osse.dc.gov/sites/default/files/dc/sites/osse/publication/attachments/Evidence-Based%20Practices%20in%20Education.pdf

Szadokierski, I., & Burns, M.K. (2008). Analogue evaluation of the effects of opportunities to respond and ratios of known items within drill rehearsal of Esperanto words. *Journal of School Psychology, 46,* 593-609.

Tan, A., & Nicholson, T. (1997). Flashcards revisited: Training poor readers to read words faster improves their comprehension. *Journal of Educational Psychology, 89,* 276-288.

Taylor, K., & Rohrer, D. (2010). The effects of interleaved practice. *Applied Cognitive Psychology, 24*(6), 837-848. https://doi.org/10.1002/acp.1598

The Iris Center. (2014). *Evidence-based practices.* Retrieved from https://iris.peabody.vanderbilt.edu/module/ebp

Tucker, J.A., (1989). *Basic flashcard technique when vocabulary is the goal.* Unpublished teaching materials, University of Tennessee at Chattanooga. Chattanooga, TN: Author

U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse.

What Works Clearinghouse, Institute of Education Sciences, U.S. Department of Education. (2022). *Growth mindset interventions for postsecondary students*. https://whatworks.ed.gov.What Works Clearinghouse. (2017). *Procedures handbook: Version 4.1.* What Works Clearinghou