Mathematics instructional level as a predictor of explicit timing fluency intervention success

Chasen Rain Dillon
University at Albany, State University of New York, crdillon@albany.edu

The University at Albany community has made this article openly available. Please share how this access benefits you.

Follow this and additional works at: https://scholarsarchive.library.albany.edu/legacy-etd

Part of the Psychology Commons

Recommended Citation
https://scholarsarchive.library.albany.edu/legacy-etd/2891

This Dissertation is brought to you for free and open access by the The Graduate School at Scholars Archive. It has been accepted for inclusion in Legacy Theses & Dissertations (2009 - 2024) by an authorized administrator of Scholars Archive. Please see Terms of Use. For more information, please contact scholarsarchive@albany.edu.
Mathematics Instructional Level as a Predictor of Explicit Timing Fluency Intervention Success

by

Chasen R. Dillon

A Dissertation
Submitted to the University at Albany, State University of New York
In Partial Fulfillment of
the Requirements for the Degree of
Doctor of Psychology

School of Education
Department of Educational and Counseling Psychology
2022
ACKNOWLEDGMENTS

This work would not have been possible without the abundance of support that I received throughout my school psychology doctoral program. I would first like to thank my committee members. This especially includes Dr. Benjamin Solomon, who served as Chairman of my Dissertation Committee. Dr. Solomon has been an invaluable contributor to my understanding of RTI and CBM which would serve as the foundation of this study’s application and relevance towards real-world scenarios within the field of school psychology. I would also like to thank Dr. Solomon for his data analysis expertise and continual guidance throughout this study. I wish to also extend thanks to Dr. David Miller for his support as a committee member, as well as Dr. Kirra Guard, for her contributions as a committee member for which she provided refreshing insight to CBM application. I would also like to acknowledge the support received from prior field experience supervisors who left a lasting impact on my evolving understanding of school psychology. This includes Dr. Mary Cloutier, Dr. Carrie Trimarchi, and Dr. Elizabeth Brown. Furthermore, I would like to acknowledge my family and friends for their continual support throughout this dissertation process as well as throughout the entirety of my school psychology program. I would like to especially thank my oldest sister, Dakota Dillon, for her guidance in navigating life and its curveballs. as well as my girlfriend, Jesslyn Chong, who lent unconditional support and insight as a fellow PHD candidate in clinical psychology.
ABSTRACT

Various evidence-based mathematic interventions have been deemed effective in remediating skill deficits. However, despite recommendation of evidence-based intervention application within Response to Intervention (RTI) frameworks, little research has been done to differentiate intervention effectiveness, specifically Explicit Timing (ET), for students who present with varying skill proficiency prior to intervention implementation. The current study utilized secondary analysis of existing ET data that was supplemented with performance feedback as part of original intervention design (i.e. ET + Feedback). The purpose was to determine whether learning rates varied among students with differential baseline proficiency with the goal of determining proficiency’s relevance to treatment effectiveness. A secondary goal was to compare the utility of assessing and monitoring proficiency with either an accuracy or fluency predictor. The sample included 77 third-grade students, who attended a public elementary school in Upstate New York. Results from Hierarchical Linear Modeling (HLM) indicated that initial proficiency is a relevant variable to consider in determining ET effectiveness, with results varying based on treatment conditions. Moreover, fluency was shown to be a superior predictor in assessing initial proficiency and monitoring treatment response. These findings are especially relevant for educators who wish to maximize treatment efforts. Future remediation efforts should consider the adoption of the following decision-making guidelines outlined in this study, to ensure that proficiency aligns with appropriate treatment, especially when considering implementation of class-wide intervention procedures for students with varying (and changing) proficiency.
<table>
<thead>
<tr>
<th>Chapter 1: Introduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purpose Statement</td>
</tr>
<tr>
<td>Possible Implications</td>
</tr>
<tr>
<td>Chapter 2: Review of Relevant Literature</td>
</tr>
<tr>
<td>Mathematic Performance in the United States</td>
</tr>
<tr>
<td>Early Numeracy</td>
</tr>
<tr>
<td>Teaching to the Common Core</td>
</tr>
<tr>
<td>Matthew Effect</td>
</tr>
<tr>
<td>Arithmetic Fluency</td>
</tr>
<tr>
<td>Response to Intervention</td>
</tr>
<tr>
<td>3-Tiered System</td>
</tr>
<tr>
<td>Instructional Match</td>
</tr>
<tr>
<td>Assessment</td>
</tr>
<tr>
<td>Need for Research</td>
</tr>
<tr>
<td>Two Decision-making Frameworks</td>
</tr>
<tr>
<td>Instructional Hierarchy</td>
</tr>
<tr>
<td>Acquisition</td>
</tr>
</tbody>
</table>
Limitations and Future Research ................................................................. 68

Chapter Summary ..................................................................................... 73

References ................................................................................................ 75
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1. Instructional Delivery Model</td>
<td>23</td>
</tr>
<tr>
<td>Table 2. Assigned 3rd Grade Instructional Level</td>
<td>37</td>
</tr>
<tr>
<td>Table 3. Instructional Group Sample Size</td>
<td>38</td>
</tr>
<tr>
<td>Table 4. Fluency Multilevel Analysis Results</td>
<td>49</td>
</tr>
<tr>
<td>Table 5. Accuracy Multilevel Analysis Results</td>
<td>52</td>
</tr>
</tbody>
</table>
# LIST OF APPENDICES

<table>
<thead>
<tr>
<th>Appendix</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appendix A. Administration Probe Exemplar: Set Size 12</td>
<td>90</td>
</tr>
<tr>
<td>Appendix B. Administration Probe Exemplar: Set Size 24</td>
<td>91</td>
</tr>
<tr>
<td>Appendix C. Explicit Timing Administration Protocol</td>
<td>92</td>
</tr>
</tbody>
</table>
Chapter 1: Introduction

Low academic achievement in mathematics along with concerns that classroom instruction alone has been insufficient in promoting the success of students meeting grade-level standards has resulted in increased implementation of intervention and preventative measures in schools. Among the measures emphasized are strategies to promote the development of arithmetic fluency skills (i.e., single-digit additional/subtraction, single-digit multiplication/division), a subtype of foundational entry-level mathematical skills which have been demonstrated to have significant implications on later mathematical achievement (Gersten et al., 2005). To ensure intervention and prevention measures are employed prior to students falling too far behind to the point where remediation is unlikely to be successful, schools have shifted towards implementation of mathematical services within a response to intervention (RTI) framework.

RTI is a data-driven decision-making framework that supports implementation of evidence-based interventions that correspond with ongoing assessment of students’ instructional needs (VanDerHeyden et al., 2007). Interventions that match students’ presented skills is of particular importance to the success of RTI and entails that specified instructional material, and the way in which that material is delivered within the intervention, aligns in difficulty with student abilities. Curriculum-based measurement (CBM) is supported by research as the preferred method of assessing and informing instructional needs within the RTI model (Hosp et al., 2016). CBM is a simple, valid, and standardized assessment measure that uses content that aligns with what students are learning in class, thereby maximizing treatment utility. However, while the benefits of CBM use in schools is becoming increasingly apparent, guidelines on how teachers can incorporate CBM mathematic measures within an RTI decision-making framework
is lacking. That is, teachers who are expected to implement CBM to facilitate an appropriate intervention-skill match are unsure how to incorporate procedures and guidelines into specific intervention selection for those students identified as at-risk.

Two heuristics presented in research that teachers might employ to facilitate instructional match are the instructional hierarchy (IH) and instructional delivery model (IDM). The IH categorizes skills into stages of developing proficiency, with each stage informing specific instructional components to optimize response to supports, thus informing intervention design (Haring et al., 1978). The IDM informs curriculum with an appropriate level of challenge by identifying a student’s instructional level (e.g., frustrational, instructional), thereby determining stage of proficiency within the IH (Gickling, 1977). With emerging research supporting the interdependent use of the IDM and IH (Codding et al., 2007), neither approach is well understood among teachers, demonstrating the challenges with transferability between research and practice. Uncertainty of which CBM metric (accuracy or fluency) to be used in assessing instructional level is an additional obstacle to teachers who intend to implement the following two heuristics. For example, Gickling (1977) suggested that instructional level be determined using accuracy scores (i.e., percent of known/unknown item), while Deno and Mirkin (1977) suggested that instructional level be determined using fluency scores, referred to as digits correct per minute (DCPM). Although an abundance of research supports the use of a fluency metric in measuring instructional level, only one study (see Burns et al., 2006) has directly compared the psychometrics properties of these measures within an applied intervention design.

Recent research has indicated the interdependent use of the IH and IDM as a vital component to maximizing student response to supports. Referred to as a skill x treatment effect, Codding et al. (2007) demonstrated that initial fluency levels predicted intervention effectiveness
as measured via student response rates. Specifically, students whose fluency fell in the frustrational range of the IDM demonstrated better response to an acquisition-stage (in reference to the IH) cover-copy-compare (CCC) intervention, while students in the instructional range of the IDM better responded to an explicit timing (ET) fluency-stage intervention. Fontenelle et al. (2020) expanded evidence of a skill x treatment effect by demonstrating differential outcomes of two fluency-stage interventions for students of different baseline instructional levels, supporting new insights that perhaps not all fluency interventions are alike for all students and that intervention effectiveness is contingent on initial instructional level. Although promising with regards to informing treatment selection, a skill x treatment effect has been relatively understudied and warrants additional replication (Fontenelle et al., 2020).

**Purpose Statement**

The purpose of the present study was to assess for additional evidence of a skill x treatment effect, demonstrating relevance of the IH and IDM in measuring instructional level and in determining which intervention to employ. Specifically, the intervention presented determined whether initial instructional level influenced student response rates to an explicit timing fluency intervention that targets arithmetic fluency. An additional purpose of this study was to provide supporting evidence for use of either a fluency or accuracy metric when assessing instructional level and monitoring treatment progress. The later was accomplished through comparing the psychometrics properties of both metrics, specifically their reliability, throughout the course of treatment.

**Possible Implications**

Evidence of a skill-treatment effect reinforces use of instructional matching in the prevention and treatment of mathematical deficits within an RTI system. The results of this study
were used to inform teachers in whether they should use students’ initial instructional level for intervention selection. Assuming initial instructional level does predict intervention effectiveness, teachers must then be equipped with guidelines and procedures to use CBM effectively. This includes using CBM within the IH and IDM to align intervention components with skill proficiency and a decision-making criterion to support whether more difficult or less difficult skills should be of target. Therefore, one implication of the presented study was to replicate mathematical research in determining whether initial instructional level is relevant to treatment response (and thus intervention selection), and to summarize, in an applied case, the procedures and guidelines teachers can use to resolve mathematical deficits, bridging the gap between research and practice.

A secondary implication of this study was to verify the measured effects of an ET fluency intervention for students of different instructional levels. While ET has been demonstrated to be an effective fluency intervention, prior research did not separate students by instructional level prior to evaluating intervention outcomes (Rhymer et al., 1998). As a result, the effects of ET may be skewed through implementation of a treatment that lacks alignment with student skill level. Measuring instructional level prior to ET implementation helped demonstrate the true effects of ET across students of varying instructional levels.

A third implication, which both assists teacher’s implementation of CBM and better explains results of prior ET research, was to clarify whether an accuracy metric or fluency metric has stronger psychometric properties and thus should be used to measure instructional level and monitor changes in outcomes. Even if teachers have a grasp with incorporating the IH and IDM, use of a non-reliable metric could lead to nonstable results and faulty instructional decision-making. Therefore, comparing the reliability of the following two metrics assists teachers in
using accurate information to inform instructional needs and in facilitating an appropriate treatment match.
Chapter 2: Review of Relevant Literature

Mathematic Performance in the United States

The 1983 publication of a 36-page report titled *A Nation at Risk: The Imperative for Educational Reform* (Gardner et al., 1983), which investigated the quality of American education, made it apparent that the educational system was insufficient, with tests scores declining and millions of students illiterate. In response, and as part of a series of legislative efforts to increase educational standards, accountability, and academic rigor, congress passed the No Child Left Behind Act of 2001 (NCLB, 2001). The NCLB Act addressed educational deficits and disparities through requiring states to increase annual testing of student achievement in mathematics and English, screen for early identification of academic difficulties, and implement evidenced-based interventions (Lee, 2010; Lee & Reeves, 2012). Among other requirements, schools were also required to participate in semi-annual national testing to serve as an independent and universal assessment measure to validate state results which often vary in content and standards (Stoneberg, 2007). The NCLB was eventually renamed as the Every Student Succeeds Act (ESSA) in 2015, but many of the same accountability measures remained intact (Mathis & Trujillo, 2016). Despite legislation’s aims to increase academic achievement across diverse student populations using a universal standards-based approach to educational reform, low academic achievement in mathematics has remained a fundamental concern for schools across the nation. In 2015, the National Assessment of Educational Progress (NAEP) reported only 41% of fourth-grade and 34% of eighth-grade students performed at or above the proficient level in mathematics (National Center for Education Statistics, 2020), suggesting that children, on a large scale, are not mastering material taught in class. Of potentially graver concern is that 2019 NAEP scores were below those reported in 2015 (National Center for
Education Statistics, 2020) reflecting a decline in performance despite years of accompanying mandates.

The intermittent decrease in standardized test scores and annual below-level proficiency performance has sparked considerable interest in determining why students are struggling in mathematics. On one end, a reactive scope of research examining the impacts of legislation blame stricter accountability efforts and insufficient resources as the catalyst for poor academic achievement in mathematics (McGuinn, 2016; Simpson et al., 2004). On the other, a proactive body of literature has emphasized the need to focus on identification and changes of classroom variables relevant to improved academic success, including expanded research on effective instructional support and interventions (Barnett et al., 2004; Bramlett et al., 2009; Gersten et al., 2009). Acknowledging which variables researchers and school personnel actually have control over, in addition to the efficacy of evidence-based intervention and prevention strategies, student-level factors, rather than legislative, have emerged as the focal point to increasing mathematical achievement. Efforts that have become especially crucial in the context of learning loss attributed to the covid-19 pandemic and school closures. Among relevant targets, early numeracy skills (i.e., foundational entry-level skills) and strategies to promote their development have been increasingly examined for treating and preventing long-term deficits in mathematics. To understand the relevance of early numeracy skills in explaining low national test scores, a review of how and under what circumstances these skills are taught, as well as implications of skill deficits, is necessary.

**Early Numeracy**

**Teaching to the Common Core.** Since its release in 2010, a large majority of states have formally adopted the Common Core State Standards (CCSS) to guide mathematical instruction
(Porter et al., 2011). Introduced as an additional accountability measure following NCLB, CCSS serves as a universal set of standards that outlines what content and skills are to be taught at each grade (i.e., grade-level standards; National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010). CCSS was developed carefully with support from the National Governors Association and the Council of Chief State School Officers (CCSSO) as well as content experts, teachers, and researchers, making it a robust set of standards considered imperative to improving scores across states with non-uniformity.

Fundamental to the CCSS and the instruction of mathematics generally is that curriculum be taught as a sequence of skills with mastery of skills at early grades serving as the prerequisite to skill development at later grades (Cobb & Jackson, 2011; Powell et al., 2013). For example, kindergarteners should focus on counting skills from 0-100 and gain conceptual knowledge of addition and subtraction. In grades one and two, students should understand how to solve addition and subtraction problems up to the value of 20, along with place values. And by grade three, students should have fully mastered addition/subtractions problems and be ready to move towards multiplication and division.

The downside, however, in creating a universal set of standards that would be considered more rigorous than previously used state standards is that while it defines expected outcomes, it does not provide schools with curriculum nor the instructional methods in which curriculum should be taught (Porter et al., 2011). This leaves teachers making idiosyncratic and often biased decisions in the selection and application of instructional methods to help students meet standards that may have once been out of reach (Fuchs et al., 1984). Meanwhile, ongoing research has continued to demonstrate teacher’s knowledge of content (i.e., what to teach) and instructional strategies (i.e., how to teach) as the most important factors to the success of student
learning in mathematics (Hill et al., 2005; Hosp & Ardoin, 2008). In summary, the outcome of higher standards and poor instructional insight poses increased risk to poorly developed foundational skills in mathematics, deficits that will unequivocally impact later skill development.

**Matthew Effect.** Considering mathematics are taught as a sequence of skills, children who fail to acquire early foundational skills will continue to struggle in courses at later grades, while the academic gap with peers will widen, in what has become coined as the “Matthew Effect” (Stanovich, 1986; Walberg & Tsai, 1983). An abundance of research has demonstrated the Matthew Effect to hold true in mathematics, demonstrating that deficits in foundational skills in early grades will lead to long term impairment to the point where children will fail to ever catch up with peers while also continuing to suffer in meeting future grade-level content standards (Chard et al., 2005; Jordan et al., 2007 Jordan et al., 2009; Gersten et al., 2005).

In a secondary analysis of 17,401 kindergarteners, Byrnes and Wasik (2009) found that pre-existing mathematic skills, in comparison of other factors believed to advance mathematic development (e.g., family socio-economic status; frequency of being exposed to mathematical content), was most strongly associated with mathematic achievement at grades kindergarten, first, and second. Furthermore, a strong foundation is not only relevant to later success in mathematics but also overall academic achievement. In a meta-analysis that included 6 distinct longitudinal studies, Duncan et al. (2007) demonstrated that among measured school-entry variables (e.g., cognition, attention, socioemotional skills), mathematic performance was the most powerful predictor of later learning. Lastly, the National Mathematics Advisory Panel (NMAP; 2008) concluded that low-income youth who acquire strong mathematical skills by the
end of middle school are much more likely to graduate from college than their peers who do not acquire such skills.

Whether it be due to environmental variables, educational differences, or instructional challenges, a percentage of students will likely fail to meet early numeracy content standards. Despite initial failure, the implications of the Mathew Effect make it seemingly clear that skill remediation occur through targeted intervention and prevention strategies to close small academic gaps as they are observed. While remediation is best accomplished when each mathematical skill is individually targeted one by one, teachers may find this skill-by-skill approach to treatment lengthy and tiresome, especially if students require remediation across multiple skills relevant to their grade level. Prioritizing which skills take precedence for remediation, mathematical research highlights fluency development as a skill with significant implications to later mathematical success (Jordan et al., 2009; NMAP, 2008)

**Arithmetic Fluency.** In efforts to remediate foundational skills deficits, much of the intervention research in mathematics has targeted the acquisition of basic arithmetic skills (i.e., learning single-digit additional/subtraction, learning single-digit multiplication/division; e.g., Gersten et al., 2005; VanDerHeyden et al., 2012). Of additional focus is mastery of fluent computation with basic arithmetic: when students recall basic facts accurately and quickly (Hasselbring et al., 1987). A reason why much of the intervention research has targeted basic arithmetic fluency is due to the predictive nature of its development. Fluency deficits are often observed in students with difficulty in mathematics, while also being a central element to those with a learning disability in mathematics (Geary et al., 2012). Students who struggle with more advanced mathematic skills presented later in the typical instructional learning sequence also often struggle with basic operational fluency (Jordan et al., 2007; Jordan et al., 2009). Moreover,
NMAP (2008) considers lacking fluency in basic arithmetic skills a fundamental concern to students’ mathematical success and identifies fluency of basic arithmetic skills as a necessary element of intervention efforts.

**Response to Intervention**

With increased emphasis on early intervention and effective methods of instructional delivery at the classroom level, many schools have begun to adopt Response to Intervention (RTI) systems (Berkeley et al., 2009). Endorsed by ESSA, RTI is conceptualized as a service delivery system that assists schools in meeting CCSS by improving academic outcomes for struggling learners before they fall behind (Lembke et al., 2012). Increased student achievement is facilitated through early identification of student needs and implementation of a continuum of evidence-based instructional supports (Gersten et al., 2009). The success of RTI is dependent on its data-driven decision-making process which provides increasingly intensive instructional support through a tiered system, evidence-based interventions that match student needs, and ongoing assessment that informs instructional success or need for change (Gersten et al., 2009; VanDerHeyden et al., 2007).

**3-Tiered System.** RTI entails a 3-tiered system of increasingly intensive instruction and intervention to ensure all students receive the support needed to meet academic standards. Having a system that provides sufficiently necessary support to advance educational outcomes of every student is fundamental to accountability efforts (Lane et al., 2013).

Tier 1 provides high-quality comprehensive instruction to all students in the general education classroom with curriculum that aligns with research supported academic content standards (Gersten et al., 2009). A strong tier 1 system is one that results in most students (75–80%) meeting defined numeracy benchmarks, with those who do not respond likely needing
more intensive tier 2 or tier 3 support (Fletcher & Vaughn, 2009). Granted that each classroom entails a variety of students with different academic abilities, it is likely that instruction may move at a pace too fast for some students to successfully master foundational numeracy skills, leading to a percentage of the student population that may require supplemental tier 2 (15%) or tier 3 (5%) support (Fuchs & Vaughn, 2012).

Tier 2 refers to additional instructional support for students who do not meet grade-level expectations at tier 1. Support at this tier should be explicit and targeted by specifying the skill-deficit and providing aligned interventions (Fletcher & Vaughn, 2009). Tier 2 interventions are often carried out in small groups (typically less than 6 students) who meet 3-5 times a week for at least 20 minutes per day over the course of 6-10 weeks (Fuchs & Vaughn, 2012).

Tier 3 services are additional supports provided to student who do not respond to tier 1 and tier 2, creating the need for more intensive intervention (Fletcher & Vaughn, 2009). Instructional support at tier 3 is typically carried out by an interventionist rather than a classroom instructor and requires more frequent and specified progress monitoring. If adequate progress following this level of support is not made, formal eligibility for special education may be considered (Vaughn & Fuchs, 2003).

**Instructional Match.** The success RTI has in remediating skill deficits through the application of tier 2 and tier 3 targeted interventions is largely determined by instructional match. Instructional match is the process of ensuring that the specified instructional material, and the way in which that material is delivered within the intervention, aligns in difficulty with the student’s abilities (i.e., proficiency; Gickling & Thompson, 1985; Haring et al., 1978). Teaching curriculum that aligns with students’ abilities is an essential element to effective learning and has been consistently found to improve student outcomes for reading (Shapiro & Ager, 1992) and
mathematics (Burns et al., 2006). That is, the material being presented is at an instructional level that the student has prerequisite skills to learn. Further, the nature of the instructional environment matches the needs of the student.

Interventions that do not align with presented proficiencies are more likely to be ineffective (as determined by student response rates), jeopardizing the integrity of RTI altogether (Burns et al., 2010; Codding et al., 2007). Ineffective academic treatment poses great risk to both current and future mathematical success, as prolonged skill inadequacy prevents students from accessing later grade-level content due to lacking prerequisite skills (i.e., Mathew Effect; Jordan et al., 2009). Furthermore, ineffective tier 2 and tier 3 supports can lead to a formal diagnosis of a learning disability in mathematics, considering unresponsiveness to RTI can be used as part of the learning disability identification process as reflected under the Individuals with Disabilities Education Act (IDEA, 2004; Kovaleski et al., 2013).

**Assessment.** To ensure instructional match, it is important to engage in assessment methods that determine student needs, which can be accomplished with two data collection approaches. Curriculum-Based Assessment (CBA) entails the collection of data from content and materials used during daily instruction to guide instructional decision, thereby also making CBA data valuable to monitor progress (Deno, 2003; Hintze et al., 2006). Progress monitoring through repeated assessments is a fundamental component to RTI, as resulting data determines whether students are making adequate progress or if changes to instruction and intervention intensity is needed (Fuchs et al., 2012; Stecker et al., 2008). CBA has become a valuable replacement to traditional norm-referenced psychometrics measures (i.e., aptitude measures) by informing instruction, as the latter approach uses assessment material differently than what is taught in the classroom, often using a summative process, resulting in poor instructional utility (Gickling &
Thompson, 1985). Because CBA uses the same materials used in instruction to test what has been learned and what needs to be learned, it is a more consistent, fairer, and provides a more adequate basis for guiding future instructional decision-making.

Curriculum-based measures (CBM) is a distinct subset of CBA that involves the incorporation of specific measurement procedures and materials used to guide the assessment process (Deno, 1985; Deno & Fuchs, 1987). In assisting in this evaluation process, it was necessary to create a simple, valid, and standardized set of measures and procedures that teachers could use to frequently measure student growth across basic academic skills (Deno, 2003). CBM measures have broad alignment with curriculum content and have been demonstrated valid across curriculums, as outlined by CCSS and state achievement tests, which supports their use in the classroom (Codding et al., 2015; Shapiro et al., 2006; Silberglitt & Hintze, 2005).

**Need for Research**

Intervention research is an important supporting element to successful RTI implementation. Similar to the content and instruction necessary to teach to the CCSS, RTI requires evidence-based resources and decision-making guidelines. However, as RTI becomes increasingly adopted in schools, the mathematical intervention research that guides its implementation is limited, which thus served as the basis for this study.

Among the lacking research is the selection of evidence-based mathematic interventions to choose from, as the majority of intervention research has predominantly addressed literacy skills (Lembke et al., 2012). While there have been increased efforts on the development of standards (Horner et al., 2005) and resources (e.g., *What Works Clearinghouse; available at ies.ed.gov/ncee/wwc*) to obtain evidence-based interventions, additional replication of their effectiveness is necessary to expand the field and validate treatment options. Therefore, one
secondary purpose of the following study was to demonstrate evidence of an effective class-wide fluency intervention. Evidence that is necessary to solidify a treatment approach for a population struggling to master fluency which has significant predictive value to later mathematical success.

Another area of research currently lacking is how reliable and valid assessment methods (i.e., CBA/CBM) can be used to ensure that interventions are appropriately selected and matched to students presented skills (Fuchs & Vaughn, 2012). RTI stresses the importance of selecting evidence-based interventions of varying intensities, matching intervention instructional components with skill proficiency, and engaging in valid and reliable assessment methods (i.e., CBM) to guide decision-making. However, guidelines on how to incorporate these measures to improve student outcomes in mathematics is lacking and warrants further research (Fontenelle et al., 2020).

**Two Decision-making Frameworks**

The following section discusses two decision-making heuristics (the instructional hierarchy and instructional delivery model) used to determine the instructional needs of students and facilitate an appropriate intervention-skill match. While both approaches have existed separately, recent studies have demonstrated significant evidence for their interdependent use in remediating and preventing mathematical difficulties (Coddington et al., 2007; Burns et al., 2010; Fontenelle et al., 2020). Nonetheless, clarified decision-making procedures, instructions, and criterion of their use is necessary. In short, the instructional hierarchy categorizes skills into stages of developing proficiency, with each stage informing specific instructional components that will optimize response to supports, thus informing intervention design. Whereas the instructional delivery model uses skill-specific assessment measures to identify student’s instructional level, ensuring that instructional content matches student abilities.
**Instructional Hierarchy**

Developed by Haring et al (1978), the instructional hierarchy is an empirically supported conceptual framework that links identified stages of academic skill development to intervention and instructional planning. Similar to how mathematic skills develop in sequential order, proficiency within each skill also develops in a hierarchical order (i.e. stages), beginning with initial exposure and progressing towards eventual mastery (Rivera & Bryant, 1992). The four linear stages of skill development include: 1) acquisition, 2) fluency, 3) generalization, and 4) adaptation. The student first acquires accuracy in a skill, becomes fluent through rapid performance, generalizes and maintains the skill across time and contexts, and lastly is able to adapt its use to changing environmental demands (Haring et al., 1978; Martens & Witt, 2004). Furthermore, because proficiency develops as a hierarchy, proficiency at one stage will support proficiency development at the next, highlighting the importance of developing each stage to satisfactory before moving to the next.

The utility of the instructional hierarchy is in its ability to increase intervention efficiency by informing the relevancy of particular treatment components in alignment with changing skill proficiency (Ardoin & Daly, 2007). Through an abundance of research focusing on skill development, Haring et al (1978) noted that children performed differently across familiar and unfamiliar tasks and that instructional procedures had varying effects on learning rates given the individual’s proficiency with the task. Results of this research led to a developed heuristic that informs instructional practices. Early research on the instructional hierarchy first emerged for improving reading skills, but 30 years of research on the instructional hierarchy has demonstrated its value as an empirically supported framework to guide components of instructional intervention across a variety of academic skills (e.g., reading, writing, mathematic;
A further explanation of each stage along with stage-dependent instructional components relevant to mathematical development is discussed below.

**Acquisition.** Acquisition is the first stage of skill development. At this stage, the individual first begins to acquire a new skill with performance typically characterized by low and variable accuracy accompanied with a slow completion rate (Haring et al., 1978). The focus at this stage is for students to produce the correct answer, without necessarily considering how long they take to produce it. Acquisition lasts from the student’s first attempt at performing the skill until they perform it with relatively high accuracy, and as such, with moderate fluency.

Enhancing accurate learning is best established when students receive instructional components of guided practice and assistance (Haring et al., 1978). Martens and Witt (2004) suggested this can be accomplished by showing the student how to perform a skill (modeling), giving them enough help to do it themselves (prompting), and then either correcting errors through immediate feedback or reinforcing correct responses. Guided practice and use of overt strategies, like those in a cover-copy-compare intervention where students reference an answer key following initial response, are often utilized and effective to facilitate accurate responding, but as accuracy improves should be reduced as these strategies are lengthy and result in dependency (Poncy et al., 2007).

**Fluency.** Students who are able to solve problems with a high degree of accuracy but remain slow in doing so exhibit fluency problems. The goal at the fluency stage is to perform an acquired skill rapidly and accurately (Haring et al., 1978). Hasselbring et al. (1987) referred to this fast and accurate recall as automaticity, which when developed for basic arithmetic facts frees up cognitive resources (e.g., attention, working memory) that can be applied towards
learning more complex tasks, such as those presented later in the skill sequence (Price et al.,
2013; Throndsen, 2011).

Fluency is best developed after the individual has developed accuracy in their response,
highlighting the linear hierarchy of the different stages of skill proficiency (Rhymer et al., 2000).
When a student becomes automatic with basic mathematic facts, which is characterized by quick,
accurate, and effortless recall, they will complete problems at a faster rate increasing
opportunities to respond to additional problems (Poncy et al., 2007). Increased opportunities to
respond, or rather repeated practice, is a necessary component to further advancing fluency
(Binder, 1996; Codding et al., 2019; Haring et al, 1978). In fact, Skinner et al. (1996) concluded
that gains in fluency are closely associated with opportunities to respond, which should be
maximized in interventions that address fluency, especially for elementary students who often
need more repetitions to retain basic math facts (Burns et al., 2015).

**Generalization.** Generalization is the third stage of skill development and entails
applying the students accurate and rapid performance to new environments different than in
those in which the skills were originally taught (Stokes & Baer, 1977; Haring et al, 1978;
Martens & Eckert, 2007). Generalization can occur across three dimensions: response, stimulus,
and time (Noell et al. 2006). Poncy et al. (2010) summarized all three types with respect to
mathematics. Response generalization entails students responding to learned stimuli using a new
behavior (e.g., solving basic mathematic problems through voice and then hand-gestures).
Stimulus generalization involves the same behavioral response but with the stimuli being
presented differently (e.g., solving horizontal number line problem through verbal response and
then responding verbally to a vertical number line). Lastly, generalization across time means a
skill has been maintained. The student can successfully complete the task at a later time (e.g., 6 weeks) along with any associated stimulus- or response- generalization as well.

Research on the generalization of various math skills has demonstrated relevance of the instructional hierarchy learning sequence. VanDerHeyden and Burns (2009) established that developed fluency promoted the generalization of skill across time and in relation to later learning. Students in grades 2-5 participated in a computational fluency-building intervention four days per week for an entire school year with retention of intervention skill measured months later. Results indicated that at all grades, children who performed with lower fluency levels (i.e., frustrational range) during the retention probe scored lower on average on each of the intervention probes than did their same-grade peers who performed with more advanced fluency scores (i.e., instructional or mastery range) on the retention probe. Additionally, it was demonstrated that children with higher performance were more easily able to master future related skills, and do so much faster, than students with lower fluency scores.

While the authors demonstrated how generalization is evidently influenced through developed fluency, the relationship between fluency and generalization is not yet definitively clear, considering that fluent responding does not always result in generalization (Hosp & Ardoin, 2008). Furthermore, the type of generalization should be evaluated. For example, Poncy et al. (2010) showed that fluency with addition facts did not generalize to related subtraction facts. In conclusion, evidence for how accurate and fluent responding with basic mathematic facts generalizes to other mathematical domains of learning is inconclusive and requires further research.

**Adaptation.** The final stage of the instructional hierarchy is adaptation. Here the skill is so automatic that the individual is able to present new variations of responses to changing
environmental demands (Martens & Witt, 2004). Adaptation has become synonymous with creativity or problem-solving, which are multifaceted and often involve unique circumstances, making measurement of adaption effects and interventions efforts scarce.

**Instructional Delivery Model**

Relevant to the application of instructional planning and intervention is ensuring that the material taught aligns with student abilities, which can be facilitated using Gickling’s (1977) instructional delivery model. The goal of this model is to provide curriculum with the appropriate level of challenge by identifying students instructional level (i.e., proficiency with a skill) or rather the “window of learning” between boredom and frustration (Gickling, 1977). The instructional delivery model includes three levels of proficiency: frustrational, instructional, and mastery (Table 1). When material is too challenging, the student resides in the frustrational level. When material is too easy, the student is considered to have mastered the material and is ready to move onto a new skill in the sequence to prevent boredom. The identification of each level can be accomplished by providing CBA measures that align with grade-level standards (e.g., basic arithmetic facts) and continuing down to less difficult skills in the hierarchy skill sequence, until each skill has an associated level. When used with CBA, performance in the frustrational level can represent the acquisition stage of skill development and performance in the instructional range represents fluency (Burns et al., 2010). This perspective is also supported with acknowledgment that NMAP supports fluency as the instructional standard (NMAP, 2008).

Identifying students’ instructional level is important as it dictates the difficulty of the instructional material presented. However, some debate exists regarding the criterion used to measure instructional level (Gickling & Thompson, 1985). Gickling (1977) suggested that the instructional level be determined based on accuracy scores, noting that the instructional level for
mathematics would consist of 70-85% known items derived as a fraction of known/unknown items. If students performed below this range, they fell in the frustrational level and if they performed above this range they fell in the mastery level. But research has since shown that accuracy alone is insufficient to differentiate instructional match due to its limitation in fully capturing skill proficiency (Lin & Kubina, 2005). Consider the example proposed by Hosp and Ardoin (2008), where two students complete the same mathematics worksheet with 95% accuracy, but with the first student completing the set number of problems in 2 minutes and the other completing the set problems in 15 minutes. Though both would be considered accurate, these students would not be comparable in performance. The first student would be considered more fluent, given they were able to solve their problems set in a shorter time period.

Deno and Mirkin (1977) suggested that the instructional level for mathematics be determined based on fluency scores (i.e., accuracy plus speed), rather than accuracy, as a fluency metric not only captures accuracy but also provides additional valuable information by including time as well (Binder, 1996). Subsequent research supports this approach as fluency data have demonstrated to be more reliable than accuracy data (Burns et al., 2006). To measure fluency Deno and Mirkin (1977) proposed a criterion that incorporates CBA measurement units of digits correct per minute (DCPM) which reflects the total number of correct numbers in their associated place value averaged across a one-minute period.

For example, a student who solves 7x7 = 48 would receive one digit correct for the 4 but not for the 8, which would then be accumulated with other correct values completed across the entirety of problems, and then averaged across one-minute intervals. Specifically, they allocated that students in grades one through three would fall in the instructional range if they scored 10-19 DCPM. Students who scored 0-9 DCPM fell in the frustrational range and students who
scored 20+ DCPM were in the mastery range. Students in grades 4-12 would use a slightly different criterion. Students were in the instructional range if they scored 20-39 DCPM. Students who performed 0-19 DCPM would be in the frustrating range and those who scored 40+ DCPM resided in the mastery range.

Continual research in the field of identifying student instructional levels has led to alternative fluency criterion. Burns et al. (2006), derived empirical evidence to support instructional level cutoffs being slightly higher than Deno and Mirkin’s (1977). In such study, instructional level was described as 14-31 DCPM for second- and third-graders and 24-49 DCPM for fourth- and fifth-graders. Additionally, various standardized web-based programs (e.g., AIMSweb and DIBELS) have been developed. Such programs compare student performance to normalized group outcomes producing individual performance as a standard score and percentile rank. For such programs, specifically AIMSweb, performance is computed with use of general outcome measures (GOM), which entails multi-skilled probes (Pearson, 2017). One limitation to these systems is that while multi-skilled probes can be helpful to assess overall grade-level content, GOMs lack specificity in identification of the specific skill a student is struggling with, reducing the utility of skill-targeted intervention (Christ et al., 2008). While at the present moment there is no definitive answer to which assessment approach is best to measure instructional level, the majority of research has utilized the criterion proposed by Deno and Mirkin (1977; Codding et al., 2007; Codding et al., 2011; Fontenelle et al., 2020), which does not rely on school’s purchase of web-based programs to increase instructional effectiveness.
Table 1

*Instructional Delivery Model*

<table>
<thead>
<tr>
<th>Stage</th>
<th>Grade</th>
<th>Accuracy Criteria</th>
<th>Fluency Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frustrational</td>
<td>1–3</td>
<td>&lt; 70% Accurate</td>
<td>0 – 9 DCPM</td>
</tr>
<tr>
<td></td>
<td>4–12</td>
<td></td>
<td>0 – 19 DCPM</td>
</tr>
<tr>
<td>Instructional</td>
<td>1–3</td>
<td>70–85% Accurate</td>
<td>10 – 19 DCPM</td>
</tr>
<tr>
<td></td>
<td>4–12</td>
<td></td>
<td>20 – 39 DCPM</td>
</tr>
<tr>
<td>Mastery</td>
<td>1–3</td>
<td>&gt; 80% Accurate</td>
<td>20 + DCPM</td>
</tr>
<tr>
<td></td>
<td>4–12</td>
<td></td>
<td>40 + DCPM</td>
</tr>
</tbody>
</table>

*Note:*  

a Developed by Gickling (1977); b Developed by Deno and Mirkin (1977)

**Assessment Criteria Reliability**

Strong psychometric properties of the assessment criteria used to measure students’ instructional level is essential to educational decision making. Reliability is one component of these properties and provides confidence that measurement results are accurate and consistent across trials. Accurate and consistent results can then be used to make informed decisions about instructional planning based on student needs (i.e., instructional match), such as the instructional level a student belongs (i.e., IDM) and the intervention components that would be of optimal benefit (i.e., IH). While a majority of research has promoted the use of a fluency criterion, little research has actually compared the reliability of accuracy and fluency criteria.

Burns et al. (2006) is one of the few who compared the alternate-form reliability coefficients of mathematic outcomes using both metrics and concluded that fluency was more reliable than accuracy. Nonetheless, additional replication is warranted. Furthermore, additional evidence may convince teachers who have grown accustomed to using an accuracy metric when grading other course related content (e.g., spelling test), to shift to fluency use when necessary and appropriate. Therefore, an additional aim of this study was to compare the reliability of
fluency and accuracy criteria to help determine which metric would more accurately inform intervention-skill match decision-making.

Evidence-based Interventions

A synthesis of research indicates there to be numerous viable mathematic interventions that have been documented to be effective at simple and moderate levels for students with varying needs of support (Codding et al., 2009). But considering research on mathematics is relatively scarce, the majority of this research has focused on acquisition and fluency (Codding et al., 2011). Two prominent interventions that been researched extensively are cover-copy-compare (CCC) and explicit timing (ET).

Cover Copy Compare. An intervention that primarily improves accuracy, and therefore with secondary gains to fluency, is CCC (Skinner et al., 1989; Poncy et al., 2007; Poncy et al., 2012). CCC involves specific procedures of providing students a worksheet of targeted problems with answers provided, but that are covered with their hand, and then asking the child to solve the problem before uncovering their answer and evaluating the accuracy of their response. If their answer is correct, they move on to the next problem. If their answer is incorrect, they are instructed to re-write their answer with the correct response.

Explicit Timing. An intervention that specifically addresses fluency is explicit timing (ET). ET is an intervention designed for improving fluency with arithmetic facts by incorporating a process of alerting students of a time limit when completing an academic task. General procedures are relatively straight forward in which students are 1) presented with a given number of mathematic facts and 2) alerted that they will have a fixed amount of time to complete as many problems as possible before the time expires. The principle behind the effectiveness of ET is that when students are made aware of a timed condition, they will
complete more problems per minute compared to non-timed conditions, encouraging fluency building (Rhymer et al., 1998; Rhymer et al., 1999; Rhymer et al., 2002). The completion of more problems per minute results in more practice and additional opportunities to respond, variables that according to the instructional hierarchy are particularly relevant to further fluency development (Haring et al., 1978).

The earliest research on ET was completed by Van Houten and Thompson (1976), who examined the effects of a timed condition on second grade students’ mathematic performance in single-digit addition and subtraction problems. In the baseline condition, students were provided a worksheet containing an assortment of basic mathematic facts and provided 30 minutes to complete their work. For the intervention phase, students were allowed the same 30-minute window to complete their work, but the teacher made it explicit that they would be timed. To increase saliency of the timed window, subjects were instructed they would be timed with a stopwatch for 1-minute timings that then repeated for 30-minutes. Results of the first trial indicated that students’ baseline rate of 5.5 problems correct per minute increased to 10.5 during the intervention phase. Equally important to the relevancy of increased response rates was that accuracy remained high throughout both trials.

Miller et al. (1995) was one of the first replication studies of Van Houten and Thompson’s (1976) study after noting a variety of relativeness weakness. Of these weaknesses, it was noted that had students been instructed to ‘work fast’ or ‘answer as many problems as you can’ that performance might have been higher during baseline when students were not timed. Secondly, it was noted that the student work period of 30 minutes for a drill and practice on simple mathematic facts was too long and that low baseline rates reported may have been attributed to students slowing down during the 30-minute work interval or quitting altogether.
Miller et al. (1995) set out to replicate the previous ET study with the goal of determining the effects on students' fluency and accuracy with mathematic facts across a series of 1-min time trials compared to baseline sessions in which students were encouraged and reminded to "work fast" and were given performance feedback. Two first grade classrooms, one general education and the other special education, participated in this study and scores were compared across two 10 minutes sessions, instead of 30 minutes. Using a multiple baseline design, researchers were able to demonstrate that during explicitly timed trials fluency scores increased, while rates decreased during baseline work periods, all without a decrease in accuracy scores, thus verifying the results of Van Houten and Thompson (1976).

As ET became more widely studied, the effects of increasing student fluency with basic arithmetic facts became apparent. Investigation of the application of ET for second grade (Rhymer et al., 1999) and third grade (Rhymer et al., 1998) elementary students confirmed a timing function to increase problem completion rates for basic mathematical facts. In both studies, when students were given basic mathematic facts during a baseline condition (e.g., no mention of a 4-minute time limit) and during the ET condition (e.g., told that they would be timed for 1-minute intervals for a total of 4 minute), problem completion rates were higher during the ET condition. One drawback, however, was that in the study by Rhymer et al. (1998) student accuracy decreased during the intervention phase.

In the two decades since ET results were first presented, additional research on ET and CBA have begun to demonstrate the invalidity of an accuracy metric as well as its low reliability, a possible explanation for results in previous studies. Nonetheless, and similar to other uncertainties regarding the field of mathematical intervention research, additional replication of intervention results, specifically ET results, is warranted.
Skill x Treatment Interaction

As deficits with acquisition and fluency of mathemetic skills become more apparent and gaps in student achievement become wider, educators have looked towards implementing evidence-based interventions to remediate deficits and improve performance. Rather than simply selecting any evidence-based intervention, educators are encouraged to engage in a decision-making process using the previously presented instructional hierarchy and instructional delivery model heuristics, to ensure that components of the intervention match presented skill stage, thereby maximizing learning outcomes. To facilitate this matching process, recent research suggests the use of fluency scores to support treatment selection across the hierarchy of mathematical skills, with evidence suggesting that fluency scores (i.e., frustrational; instructional) can predict the effectiveness of the intervention depending on its instructional components, thus bridging the instructional hierarchy with the instructional delivery model (Codding et al., 2007; Burns et al., 2010, Fontenelle et al., 2020). This approach of using fluency scores to inform intervention selection is referred to as a skill x treatment interaction, and although relatively new in the professional literature, is gaining greater relevance with informing effective instructional design.

Codding et al. (2007) were the first to draw explicit evidence for a skill x treatment interaction, before the term was used, by investigating whether initial fluency levels predicted intervention effectiveness. In their study, 98 second and third grade students were randomly assigned to one of three groups: CCC, ET, or a control group. The goal was to then determine if the following treatment conditions were equally effective for students' whose initial fluency levels were in the frustrational (<10 DCPM) and instructional (10 to 19 DCPM) ranges. Although both CCC and ET have demonstrated to increase fluency (Poncy et al., 2007; Poney et
al., 2012), the researchers hypothesized that students in the frustrational range would benefit most (i.e., produce the greatest increase in slope over time) from CCC, whereas students in the instructional range would benefit most from ET. This was because each intervention has instructional components that make it more suitable for one type of learner than the other based on initial fluency proficiency (i.e., stage within the instructional hierarchy). CCC was hypothesized to be more suitable for students in the frustrational range (i.e., acquisition stage) because it incorporates modeling, prompting, response feedback, and overt strategies, whereas ET might better support those in the instructional range (i.e., fluency stage) as a result of increased opportunities to respond. Frustrational students in either intervention group who did not respond with satisfactory growth would be recommended to receive an intervention on a prerequisite skill in the learning sequence.

The following hypothesis was confirmed in which students whose initial fluency level fell in the instructional range and who participated in the ET condition displayed higher growth in performance at the end of the intervention period than children in the other two conditions (Codding et al., 2007). Meanwhile, children who initially performed in the frustrational range and participated in the ET group had the lowest growth in performance compared to the other groups with progress also being slower over time. The authors noted no differences in performance for the CCC group compared to the control group. Results suggested that not all fluency problems are equivalent and that different interventions may improve rates of responding depending on whether initial fluency falls in the instructional or frustrational ranges (Shapiro, 2004).

Further evidence for a skill x treatment interaction was found by Burns et al. (2010), who conducted a meta-analysis on mathematic intervention research utilizing CBA and CBM in
combination with the instructional delivery model and instructional hierarchy. Their goal was to use meta-analytic procedures to examine the effectiveness of acquisition and fluency interventions according to baseline performance that fell within the frustrational and instructional levels. Among 1011 articles reviewed, 17 peer-reviewed studies met their inclusion criteria, with the most prominent criteria being that 1) DCPM data was collected pre- and post- intervention, allowing for fluency growth rates to be measured, and 2) the intervention utilized was described with enough detail to determine whether it was an acquisition of fluency intervention.

A total of 11 studies used acquisition interventions, 5 used a fluency intervention, and one study combined an acquisition and fluency intervention. Baseline data across the 17 studies, which was used to measure students’ instructional level, thus determining appropriateness of intervention selection (e.g., fluency or acquisition), relied on the fluency and accuracy criteria set forth by Burns et al. (2006). Students were identified as being in the instructional level if their score fell within 14-31 DCPM for second- and third-graders and 24-49 DCPM for fourth- through sixth-grade students. Scores that fell below the lowest end of the respective ranges (14 and 24 DCPM) were identified as frustration-level skill. Scores that fell above 31 DCPM for second and third grade and above 49 DCPM for fourth- through sixth-grade were in the mastery level.

Results from this meta-analysis suggested sufficient support for a skill x treatment interaction. Students who initially tested in the frustrational range (signifying the acquisition stage of the instructional hierarchy) at the start of the intervention benefited from larger growth when they received an acquisition intervention (effect size .84), compared to those who tested in the instructional range and received an acquisition intervention who displayed only moderate effects (.49). Effect size was calculated using students’ percentage of nonoverlapping data (PND;
Scruggs et al., 1987), a nonparametric effect size that is based on Cohen’s (1988) standardized mean difference criteria and that can be computed by counting the number of intervention data points that exceeded the highest baseline data point and dividing by the total number of intervention points.

Conclusions on the effects of the fluency intervention across both frustrational and instructional groups was inconclusive given only a small sample of students were initially measured to be in the instructional range, highlighting the issue that many students are struggling to gain fluency in typical school curriculum and highlighting the importance of furthering fluency research. However, the authors did note that fluency interventions had only a small to moderate effect (.47) on students performing at a frustrational level, which was approximately the same magnitude as the effect for acquisition interventions on students performing at the instructional level (.49), furthering partial evidence for a skill x treatment effect.

More recently, Fontenelle et al. (2020) demonstrated evidence for a skill x treatment effect by comparing performance growth of students receiving one of two class-wide fluency interventions. This study investigated whether initial fluency scores would differentiate intervention effectiveness between Taped Problems and ET procedures. Taped Problems (TP; Poncy et al., 2007) is a fluency intervention implemented by providing students with a worksheet containing basic computation problems. The teacher plays an audio file that reads each mathematic problem, provides a brief pause, and then states the answer. Students are instructed to try and beat the pause in the audio file and write the correct response to each problem before the answer is heard. If the student fails to respond in the allotted time or writes down the wrong answer, they cross out the response and write the correct answer. The TP intervention
incorporates brisk pacing to increase rates of responding and includes an immediate feedback component as the answer to each question is provided (Fontenelle et al., 2020).

Results showed that both TP and ET resulted in similar DCPM increases, with both approaches significantly outperforming a control condition. The interaction between intervention type and initial fluency levels suggested a differential effect. For students with initial fluency scores in the frustrational range (≤10 DCPM), it was found that TP resulted in a higher growth rate than both ET and the control. For students with initial fluency scores in the instructional range (>10 DCPM), both ET and TP interventions resulted in higher DCPM scores than control but did not differ from one another. Similar to prior comparisons of CCC and ET (Codding et al., 2007), while both ET and TP interventions address fluency deficits, instructional components may differentiate their effectiveness based on student proficiency within the instructional hierarchy.

TP was expected to better support learners in the frustrational range because of the audio pause and immediate feedback, which are needed to promote both accurate and fluent responding for acquisition learners. ET, on the other hand, removes these components, which results in higher response rates which has been shown to be effective for students in the instructional range but may remove components helpful to students in the frustrational range. An additional consideration for use of either intervention is whether instructional content needs to be differentiated to students in class-wide groups and with whom skill proficiency varies. Explicit timing has the advantage in which content can be specific to each student whereas TP assumes all students are working on the same skill and at the same pace, potentially slowing the learning process for students with more advance fluency as opportunities to respond are reduced to meet class-wide needs.
Chapter Summary

National assessments of mathematics have demonstrated U.S. scores to be annually below average with some evidence of decline. Below standard performance, despite increased legislation, has shifted treatment towards student-level intervention and prevention efforts. Among the efforts is increased research on effective strategies to promote students’ development of mathematical fluency, which in the context of early foundational skills has primarily pertained to arithmetic skills. RTI has emerged at the primary system with which prevention and intervention measures are applied; however, implementation guidelines for teachers are lacking. This includes teachers’ abilities to select and implement evidence-based interventions of varying intensities, match intervention instructional components with skill proficiency, and engage in valid and reliable assessment methods (i.e., CBM) to guide decision-making.

The instructional hierarchy (IH) and instructional delivery model (IDM) have become two prominent decision-making heuristics to help teachers determine students’ instructional needs and facilitate an appropriate intervention-skill match. The IH categorizes skills into stages of developing proficiency, with each stage informing specific instructional components that will optimize response to supports, thus informing intervention design. The IDM uses skill-specific assessment measures to identify student’s instructional level, ensuring that instructional content matches student abilities. While both heuristics have existed separately in research, applied cases of their interdependent use is limited but relevant to successful treatment. Nonetheless, if it is expected that teachers use the following heuristics to resolve mathematical deficits, then they must be equipped with clear instruction of how to measure instructional level (using the IDM) and how to select an intervention with appropriate components (using the IH). Necessary, but also limited in research, is evidence of the appropriate metric (accuracy/fluency) to use within
these decision-making processes, revealing the additional need to address the reliability of both measures.

The collective use of the IH and IDM in treatment of mathematical skills has suggested the existence of a skill x treatment effect. A skill x treatment effect suggests that students’ initial instructional level (i.e., frustrational, instructional) can predict intervention effectiveness as measured via progress growth. Furthermore, it implies that not all interventions are equal and that some treatments may have instructional components that are better suited for students based on initial instructional level. For example, while both cover-copy-compare and explicit timing are considered to be effective fluency interventions, cover-copy-compare may better support those in the frustrational stage and explicit timing may better support those in the instructional stage.

**Research Questions**

The study sought to verify existence of a skill x treatment effect and relevance of a fluency metric in decision-making processes, and as a result addressed the following research questions:

1. Does student initial instructional level, measured as either accuracy or fluency benchmarks, predict differential treatment response to an explicit timing fluency intervention that is supplemented with performance feedback?

   a. If one or both metrics (accuracy and fluency) predict intervention success, which one is the better predictor?

2. Which metric, accuracy or fluency, is the more reliable metric for quantifying student performance over time?
Chapter 3: Methodology

Overview

This chapter describes the methods and procedures used to evaluate differences in the effectiveness of a mathematics intervention for students of various baseline instructional levels. A description of analytical design, participants, measures, and procedures to determine intervention and criterion effectiveness follows.

Design

This study was a secondary analysis of within-person experimental mathematics data collected in January 2020, when four third grade classrooms received a four-week class-wide explicit timing intervention, which was supplemented with a performance feedback component. Feedback, in the form of error correction, notified students of problems answered incorrectly, with the correct answer then listed adjacent. Although not a component of traditional ET procedures, error correction was implemented as a supplemental component of the intervention to maximize performance growths by ensuring that students are not routinely practicing incorrect answers, a process which has been shown to benefit accurate and fluent responding (Rapp et al., 2012, Rhymer et al., 2000). Aside from the inclusion of error correction, the remaining ET guidelines and procedures followed those outlined in prior ET literature (e.g., Van Houten & Thompson, 1976; Rhymer et al., 2002). The intervention with which data was collected is described in an unpublished manuscript by Payne et al. (2021) and will be summarized here. Intervention procedures occurred three times per week and assessment twice per week producing a total of 12 intervention data points and 8 assessment data points. Performance was tracked over time as fluency (i.e., DCPM). As an independent contribution to this study and to answer the second research question of determining whether an accuracy metric or fluency metric is more
reliable across time, class-wide data collected as part of Payne’s et al. (2021) original study was rescored for accuracy across participants and conditions.

**Set Size.** Set size was one of two conditions imposed as part of original intervention design. Within each of the four classrooms, students were randomly assigned to either set size 12 \((n = 38)\) or set size 24 \((n = 39)\) condition. Set size reflects the number of unique problems targeted on a specified probe, in contrast to total problems presented (Poncy et al., 2015). For example, a single probe with 50 problems and set size 5 would entail 5 unique problems repeated in a non-specified order ten times over. Set size is typically determined by objectives specific to each intervention, although differences in set size have been considered to influence probe difficulty and precision when estimating growth (i.e., reliability; Solomon et al., 2019). Precision has direct implications for decision-making errors in which probes with lower precision will require a longer duration of progress monitoring to achieve a reliable estimate of slope (Christ, 2006). While set size was not a direct aim of this study, due to differences in psychometrics properties of probes with different set size and recommendations that intervention effectiveness not be evaluated across probes of different set size (see Solomon et al., 2019), the two set size conditions were analyzed separately. That is, although they reflect the same content, they were treated as separate outcome variables. Thus, in measuring differential treatment growth across instructional groups, students were separately evaluated across set size 12 and set size 24 conditions.

**Goal Setting.** Goal setting was the second condition introduced as part of the original intervention design in Payne et al. (2021). In assigning this condition, two of the four participating classrooms were randomly selected to receive goal setting, with the remaining two classrooms being assigned to no goal setting. While set size was randomized at the student level,
goal setting was randomized by classroom because the goal setting directions were verbally stated to students each session. Goal setting refers to whether students were directed to individually graph their intervention progress, which provided visual feedback of performance, a variable potentially relevant to increasing motivation and thus performance (Gross et al., 2014). Although goal setting was an independent variable, this condition was ignored in the secondary analysis on the basis of Payne’s et al (2021) results, which demonstrated that the inclusion of goal lines on graphs promoted no significant improvement with basic mathematics skills. That is, no significant effects between the goal setting and no goal setting groups was found. We therefore assume equivalence across levels of this condition.

**Instructional Level.** Instructional level is a condition not made explicit as part of the original intervention design in Payne et al. (2021) but was added as a key component to this study. Using both fluency and accuracy criteria, instructional level was assessed across participants at Time 1 to compare differences in performance throughout the intervention period. Instructional level based on fluency scores used Deno and Mirkin’s (1977) criterion, while instructional level based on accuracy scores used Gickling’s (1977) criterion. Baseline instructional levels was determined by comparing students’ score on the first initial assessment probe in accordance with the described third grade fluency criterion or third grade accuracy criterion (see Table 2). Furthermore, to retain a moderate sample size when comparing performance across instructional groups, students who performed at the ceiling during the initial assessment period were categorized within the instructional level group. In other words, students who tested in the mastery level were reassigned to the instructional level resulting in a slightly modified accuracy and fluency grouping criteria. Interpretation of these effects are further reviewed in the discussion of this study.
Table 2

Assigned 3rd Grade Instructional Level

<table>
<thead>
<tr>
<th>Instructional Level</th>
<th>Accuracy Criteriaa</th>
<th>Fluency Criteriab</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frustrational</td>
<td>&lt; 70 % Accurate</td>
<td>0 – 9 DCPM</td>
</tr>
<tr>
<td>Instructional</td>
<td>70 – 100 % Accurate</td>
<td>10 + DCPM</td>
</tr>
</tbody>
</table>

*Note: aAdapted from Gickling (1977); bAdapted from Deno and Mirkin (1977)*

Participants

Participants included 77 third grade students, all of whom attended a public elementary school located in Upstate New York at the time of the Payne et al. (2021) study. Participants originated from four separate third grade classrooms; three general education classrooms and one inclusive classroom, with the latter including a portion of students identified as receiving special education services that are more integrated into class curriculum than in general education classrooms. Demographics of participating students were not specifically measured as part of data collection procedures, although publicly available district-level demographics reported a population comprised of 47% White, 17% Black, 17% Latinx, 8% Asian, and 11% Multiracial; 47% female and 53% male; 71% of students who were economically disadvantaged based on qualifications to free and reduced lunch; and 6% English Language Learners. All 77 students remained as participants for the entirety of the four week in-person intervention with no student attrition. However, based on archival data collection procedures only 38 of 77 students had accuracy data collected, resulting in a difference in sample size across conditions. The adequacy of sample size as well as differences in sample distribution across conditions are later addressed as possible limitations.

Identification of instructional level was determined by comparing student’s performance on their first assessment probe to Deno and Mirkin’s (1977) fluency criterion as well as
Gickling’s (1977) accuracy criterion, with students at the mastery level assigned to the instructional level group. The sample size of each instructional group, as dictated by the grading criterion used and with consideration of the set size condition is reflected in Table 3. Comparison of initial assessment probe performance to instructional level using Deno and Mirkin’s (1977) fluency criterion indicated that, of the 77 participants, 22 students performed in the frustrational level and 55 performed in the instructional level. Using Gickling’s (1977) accuracy criterion, an analysis of the 38 sampled participants indicated that 9 performed in the frustrational level and 29 performed in the instructional level (Table 3). The number of students who initially measured in the mastery level but were re-categorized to the instructional level included 19 students using a fluency criterion and 26 students using an accuracy criterion.

Table 3

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Set Size</th>
<th>Instructional Level</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (i.e., % total problems correct)</td>
<td>12</td>
<td>Frustrational</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>Instructional</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>Frustrational</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>Instructional</td>
<td>16</td>
</tr>
<tr>
<td>Fluency (i.e., DCPM)</td>
<td>12</td>
<td>Frustrational</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>Instructional</td>
<td>26</td>
</tr>
</tbody>
</table>

Note: Instructional level of total participants assessed using accuracy metric and fluency metric.

Primary Outcome

Target Skill. Single-digit multiplication was selected as the targeted skill within the ET intervention on the basis of NY mathematics standards, classroom curriculum, and teacher input. New York State Common Core Learning Standards for Mathematics outlines that third-grade
instruction focus on developing understanding of multiplication and strategies for multiplication within 100 (Engage NY, 2015). Course curriculum, which entailed a K-5 mathematics program endorsed throughout New York State titled \textit{Go Math!}, supported a similar goal, noting that by the end of Grade 3 students be expected to know, from memory, all products of two, one-digit numbers (Houghton Mifflin Harcourt, 2015). Furthermore, all four classroom teachers supported single-digit multiplication as an appropriately targeted skill reflective of student needs.

\textbf{Application.} Single-digit CBM probes were created to accommodate both assessment and intervention processes. To account for the four-week intervention period, a set of 8 assessment probes and a set of 12 intervention probes were created for both set size 12 and set size 24 conditions. The initial assessment probe was used to measure baseline performance, assigning students to frustrational or instructional groups (in the present study), with the remaining being used to monitor changes in performance throughout the intervention. Conversely, the intervention probes were used as instructional materials for intervention and their administration did not follow standard CBM conventions, and therefore were not analyzed.

\textbf{Content.} All probes within each condition included a random assortment of single-digit multiplication problems. Randomization was generated using Microsoft Excel. Possible problem types included those with number multipliers ranging from x2 to x9, and individual multipliers were selected for each condition randomly without replacement from the available pool of problems. Multipliers of x0, x1, and x10 were excluded to remove problems considered too simple and unlikely of equal difficulty as to the other problems. Probes contained 64 problems oriented across eight columns and eight rows on a single-sided sheet of portrait-paper. Intervention probes were double-sided, adding an additional 64 problems in the event students finished their problem set early.
Each probe within its associated set size condition randomly generated either 12 or 24 unique problems that would repeat in a block-randomized manner accounting for the 64-total problems (see Appendix A and B). For example, for each set size 12 probe, each of the 12 randomly assigned problems was included once in a random order in the first block of 12 problems on the page and again for the second block of 12 problems on the page, until a total of 64 problems were generated on the single-sided paper. This process was repeated across all probes. Block-randomized stratification of items was utilized to ensure consistency in the difficulty of alternative form assessment probes, a process that has been demonstrated to lead to more reliable progress monitoring in comparison to random item arrangement (Methe et al. 2015). This is because random item arrangement may present the same problem in successive order or within close proximity on some probes but not others, creating inequivalent probes. Furthermore, to ensure maximal randomization, probes were also manually reviewed and modified so that the same problem was not repeated consecutively and did not appear directly above or below the same unique problem on the page.

**Procedures**

Intervention occurred three days per week for four weeks in a school setting. Assessment probes and associated procedures were administered Monday and Friday of each week and intervention probes and associated procedures administered Monday, Wednesday, and Friday. On days when assessment and intervention probes were administered together, assessment probes were administered first to prevent carry-over effects. Each student was assigned a folder that contained each session’s associated intervention and assessment probe. Folders were distributed at the start of each session and collected at the end. To ensure students completed the correct probe at the appropriate time, assessment probes were filed on the left side pocket of the
folder and intervention probes on the right. Task directions then specified which probe to complete at the appropriate time.

Also included on the left pocket, behind the assessment probe, were students’ prior session’s intervention probe. Marked with red pen, these probes were error corrected between intervention sessions by the researchers, with incorrect answers circled and correct answers penciled in. Error correction is simply corrective feedback to ensure that students are not routinely practicing incorrect answers and has been shown to benefit accurate and fluent responding (Rapp et al., 2012, Rhymer et al., 2000). Error correction was implemented as an additional component to the intervention to maximize performance growths. Although more common in the acquisition stage of the instructional hierarchy, feedback was summative across sessions, a sharp distinction to acquisition-based interventions which provide corrections immediately following presented items.

Assessment. Assessment periods began when students were asked to open their folder and retrieve their “math worksheet” from the left side of their folder (i.e., assessment probe). After students closed their folders and placed them aside, the interventionist then read aloud assessment procedures verbatim (see Appendix C). Assessment procedures informed students they would have two minutes to complete as many problems that they could, starting with problem one in the upper left corner of the page and working their way across and then down. All timing processes were measured using a stopwatch. To prevent students from spending excessive time on any one problem, thereby confounding assessed performance, students were instructed that if they did not know the answer to a problem they should place an “X” through the item and move on to the next item in the series. Throughout the two-minutes, the interventionist closely monitored on-task behavior. At the conclusion of two minutes, students
were instructed to put their pencil down and put their math worksheet back in the left side of their folder.

**Intervention.** Intervention periods followed similar procedures as assessment, with one exception being that students first were provided access to their scored prior day’s intervention probe, so as to promote accuracy. Placed behind the assessment probe on the left side of the folder was student’s prior intervention probe with error corrections. As part of the error correction procedures, students were directed to spend one-minute reviewing their probe while whispering the correct answer to themselves. After the one-minute period, students placed the probe back in the left side of their folder. Next, they were asked to retrieve their “math worksheet” from the right side of their folder (i.e., intervention probe). Students were instructed they would have four minutes to complete as many problems as they can, again starting in the upper-left corner of the page and working their way across and down, and then to continue solving problems on the back of their worksheet if they finished the front page. After four minutes elapsed, the interventionist asked students to put their pencil down and place their worksheet back in the right pocket of their folder, thereby concluding that day’s sessions.

Data collection and scoring were completed by graduate-level school psychology students. To support implementation fidelity, data collectors were provided a 1-hour training on correct intervention administration and CBM scoring prior to the start of the four-week intervention. Furthermore, for each session, interventionists had access to the intervention protocol which outlined intervention procedures, a script of intervention directions to be read verbatim to students, and explicit directions to monitor student’s on-task behavior (see Appendix C). Intervention treatment integrity was also assessed using checklists that were administered to classroom teachers at the beginning of each session. Teachers were asked to complete this
checklist during each session to ensure that data collectors appropriately followed all specified intervention procedures outlined in the protocol.

**Data Analysis**

This study utilized hierarchical linear modeling (HLM; Raudenbush et al., 2011) to examine the relationships between individual outcomes over time and associated initial instructional level grouping. HLM is a beneficial means to analyze changes over time when a series of individual repeated measures (Level 1; e.g., fluency data, accuracy data) are nested within differential groups (e.g., frustrational, instructional); in this case, data being nested within students (Level 2). The benefit of HLM over traditional linear regression is that it reduces the inflation of type I error by basing its sample size on the number of individuals rather than the total number of observations and also separates within-group effects and variance from between-group effects and variance. HLM also permits consideration of additional model outputs, such as variability of effects across student clusters. Failure to account for nested data structures generally leads to misinterpretation of results which can be attributed to aggregation bias, misestimated standard errors, and heterogeneity of regression (Raudenbush & Bryk, 2002).

To answer research question 1, the independent variables included a quasi-experimental grouping variable of initial fluency level. The two predictors, which were used to determine instructional level grouping, and which were separately assessed across set size 12 and set size 24 conditions, comprised of Deno and Mirkin’s (1977) fluency criterion and Gickling’s (1977) accuracy criterion. Criteria were slightly modified to account for students measured in the mastery level who were regrouped to the instructional level (i.e., instructional and mastery groups were collapsed). Each independent variable included two levels. The two levels of the fluency metric predictor were defined as either initial frustrational level fluency (DCPM ≤ 10) or
initial instructional level fluency (DCPM > 10). The two levels of the accuracy metric predictor were defined as either initial frustrational level fluency (accuracy ≤ 70%) or initial instructional level fluency (accuracy > 70%). The dependent variables included student DCPM scores on the independent sets of math probes, which are assumed to have equal variance, and which align with their associated predictor’s metrics.

In this model, time/growth rate was a level 1 variable (i.e., 1, 2, ..., 8), and initial performance (fluency; accuracy), which influences the growth rate, was a level 2 variable. The level 2 variables were represented by dummy codes (0=low fluency, 1=instructional fluency). The cross-level interaction between the level 1 predictor and instructional fluency level status was reviewed to examine the degree that instructional level status has on the relationship between explicit timing intervention trials and change in outcome over time. A skill-by-treatment interaction was to be observed if one or both metrics (initial fluency or accuracy) predict future growth under intervention. The degree to which an accuracy metric or fluency metric better predicted intervention success was determined by comparing the significance of associated p-values.

HLM was also used to answer the second research question. Here, accuracy or fluency performance was positioned as the outcome and a level 1 fixed effect for time was again added to the model. The two models (outcome of accuracy or fluency) were compared using indices of model fit, as well as inspection of reliability estimates, which were calculated as the proportion of within-student variance divided by this value and error variance (Raudenbush & Bryk, 2002). Descriptives were also calculated for growth over time when measured using both metrics.
Chapter 4: Results

Nested datasets do not necessarily require multilevel modeling. If there is no variation in response scores across level-2 units (e.g., students), the data can be analyzed using ordinary least squares (OLS) multiple regression. However, if there is variation in the mean outcome levels across students (greater than 5%), multilevel modeling is needed to separately estimate the outcome variance that occurs both within individual repeated measures and across groups (Bliese, 2000).

A preliminary calculation of the proportion of variance explained by the grouping structure in the population was used to determine the utility of HLM. Calculated intraclass correlation (ICC) ranges from 0 to 1.0 and describes the proportion of the total variance that depends upon group membership (Hox, 2010). ICC results indicated significant variability in the data as a result of person-level influences across both the fluency (DCPM) predictor and accuracy predictor as well as across set size conditions. The ICC for the fluency predictor across set size 12, ICC = .66, and set size 24, ICC = .68, reflected that 66% and 68% of the variance in DCPM scores respectively depended upon the student. The ICC for the accuracy predictor across set size 12, ICC = .62, and set size 24, ICC = .63, reflected that 62% and 63% of the variance in accuracy scores respectively depended upon the student. Therefore, results were indicative of appropriate HLM need.

Level 1 Model

The level 1 model (i.e., within-person model) measured individual student initial outcome scores and associated change over time (Singer & Willett, 2003). In this case, the following level-1 model reflected a linear regression of individual student outcomes (i.e., using either DCPM or accuracy) across intervention sessions:
The equation for DCPM is given by:

$$DCPM_{it} = \pi_{0i} + \pi_{1i} \times (TIME_{it}) + e_{it}.$$ 

At level 1, $DCPM_{it}$ represented the outcome $Y$ for level one unit $t$ (session) for the $ith$ (person), and was equal to the level one intercept, $\pi_{0i}$, the specific observation which is assumed to increase linearly over time ($TIME$), $\pi_{1i} \times (TIME_{it})$, and residual or unexplained variance $e_{it}$ that reflected the difference between each student's observed and predicted DCPM scores. $TIME$ simply represented the time between testing occasions (i.e., level-1 predictor).

**Level 2 Model**

Level two of the HLM growth model (between-person model) took each student slope and intercept coefficient from level 1 and defined its own regression equation, while accounting for grouping. A binary level-2 dummy predictor variable (ACQ), frustrational group (0) or instructional group (1), was added to the Level-2 model to explain intercept and slope variance for each associated outcome variable (i.e., DCPM or accuracy).

$$\pi_{0i} = \beta_{00} + \beta_{01} \times (ACQ_{i}) + r_{0i}$$

$$\pi_{1i} = \beta_{10} + \beta_{11} \times (ACQ_{i}) + r_{1i}.$$ 

Adding instructional group as a binary dummy predictor variable, the intercept ($\beta_{00}$) was interpreted as the average DCPM score for frustrational students (i.e., the group coded zero), and the slope ($\beta_{10}$) is the average change in frustrational students' DCPM scores per assessment period increase. Consistent with the use of dummy code predictors in multiple regression analyses, the ($\beta_{01}$) coefficient is the mean instructional group difference at time period zero, and
the coefficient ($\beta_{11}$) is the average difference in slope between groups. Also included is a residual term that reflects individual student differences in outcome about the grand mean ($r_{i1}$).

**Mixed Model**

The mixed model – a rewriting of the level 1 and level 2 equations into a single equation - simultaneously models fixed effects, which are average sample-level trends (level-1), as well as random effects (level-2), which are the extent to which these trends vary across groups. The final mixed model equation is reflected by the following:

$$DCPM_{it} = \beta_{00} + \beta_{01} \cdot ACQ_i + \beta_{10} \cdot TIME_{it} + \beta_{11} \cdot ACQ_i \cdot TIME_{it} + r_{0i} + r_{1i} \cdot TIME_{it} + \epsilon_t.$$  

**Fluency Outcome**

**Set Size 12.** Results showed a non-significant mean DCPM score for frustrational students at time zero, $\beta_{00} = 1.17; t(36) = .83, p > .05$, while instructional-level students had significantly higher starting DCPM scores at time zero, $\beta_{01} = 14.57; t(36) = 6.71, p < .01$, compared to the frustrational group (Table 4). Additionally, students in the frustrational group showed significant gains in DCPM per intervention session, $\beta_{10} = 2.84, t(36) = 7.48, p < .01$, while instructional-level students also showed gains and which were significantly greater than students in the frustrational group, $\beta_{11} = 1.62, t(36) = 2.78, p < .01$.

Level-1 variance of intercepts and slopes across subjects was also measured. The results indicated that the variance of the intercepts, 37.91, was significantly greater than 0, $\chi^2 (36) = 282.08, p < .01$, suggesting that the individual intercepts differ more than would be expected on the basis of chance. The variance in the slopes, 2.84, was also significantly greater than 0, $\chi^2 (36)$
\[ 396.75, \ p < .01, \] suggesting that there are true differences in individuals’ slopes. Variance attributed to error was 8.19.

**Set Size 24.** For set size 24, results showed a non-significant mean DCPM score for frustrational students at time zero, \( \beta_{00} = -0.39, \ t(37) = -0.13, \ p > .05 \), while instructional level students had significantly higher starting DCPM scores at time zero, \( \beta_{01} = 20.57, \ t(37) = 5.54, \ p < .01 \), compared to the frustrational group (Table 4). Additionally, students in the frustrational group showed significant gains in DCPM per intervention session, \( \beta_{10} = 4.64, \ t(37) = 8.33, \ p < .01 \), while instructional students also showed gains but these which were less than and marginally significantly different than students in the frustrational group, \( \beta_{11} = -1.35, \ t(37) = -1.97, \ p = .05 \).

Level-1 variance indicated that the variance of the intercepts, 110.87, was significantly greater than 0, \( \chi^2 (36) = 538.23, \ p < .01 \), suggesting that the individual intercepts differ more than would be expected on the basis of chance. The variance in the slopes, 3.55, was also significantly greater than 0, \( \chi^2 (36) = 303.73 \ p < .01 \), suggesting that there are significant differences in individuals’ slopes. The variance due to error was 12.59.

Overall, results indicated that students measured in the instructional group at the beginning of the intervention phase had significantly higher DCPM scores. Average scores for students in the frustrational group had starting values ranging from 1.17 DCPM for set size 12 to 0 DCPM (rounded to 0 from -.39) for set size 24, with average scores for students in the instructional group ranging from 15.74 DCPM for set size 12 to 20.18 DCPM for set size 24. Regarding changes in scores throughout the intervention period, both groups displayed gains in DCPM scores. For set size 12, students in the frustrational group showed a significant average increase of 2.84 DCPM per intervention session, with students in the instructional group having an average growth rate of 4.46 DCPM. For set size 24, growth rates were comparable, with
frustrational students increasing at a rate of 4.64 DCPM and instructional students growing at a slightly smaller rate of 3.29 DCPM. An explanation for differences in group response rates across set size is further addressed in the discussion section.

Table 4

Fluency Multilevel Analysis Results

<table>
<thead>
<tr>
<th>Set Size 12</th>
<th>Fixed Effects</th>
<th>Parameter</th>
<th>Coefficient</th>
<th>SE</th>
<th>t</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>( \beta_{00} )</td>
<td>1.17</td>
<td>1.41</td>
<td>0.83</td>
<td>36</td>
<td>.415</td>
<td></td>
</tr>
<tr>
<td>INSTR</td>
<td>( \beta_{01} )</td>
<td>14.57</td>
<td>2.17</td>
<td>6.71</td>
<td>36</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>Slope</td>
<td>( \beta_{10} )</td>
<td>2.84</td>
<td>0.38</td>
<td>7.48</td>
<td>36</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>INSTR</td>
<td>( \beta_{11} )</td>
<td>1.62</td>
<td>0.58</td>
<td>2.78</td>
<td>36</td>
<td>.009</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variance Components</th>
<th>Parameter</th>
<th>SD</th>
<th>Variance</th>
<th>df</th>
<th>( \chi^2 )</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>( r_0 )</td>
<td>6.16</td>
<td>37.91</td>
<td>36</td>
<td>282.08</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>TIME Slope</td>
<td>( r_1 )</td>
<td>1.69</td>
<td>2.84</td>
<td>36</td>
<td>396.75</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Level-1 Error</td>
<td>e</td>
<td>2.86</td>
<td>8.19</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Set Size 24</th>
<th>Fixed Effects</th>
<th>Parameter</th>
<th>Coefficient</th>
<th>SE</th>
<th>t</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>( \beta_{00} )</td>
<td>-0.39</td>
<td>3.03</td>
<td>-0.13</td>
<td>37</td>
<td>.899</td>
<td></td>
</tr>
<tr>
<td>INSTR</td>
<td>( \beta_{01} )</td>
<td>20.57</td>
<td>3.72</td>
<td>5.54</td>
<td>37</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>Slope</td>
<td>( \beta_{10} )</td>
<td>4.64</td>
<td>0.56</td>
<td>8.33</td>
<td>37</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>INSTR</td>
<td>( \beta_{11} )</td>
<td>-1.35</td>
<td>0.68</td>
<td>-1.97</td>
<td>37</td>
<td>.054</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variance Components</th>
<th>Parameter</th>
<th>SD</th>
<th>Variance</th>
<th>df</th>
<th>( \chi^2 )</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>( r_0 )</td>
<td>10.53</td>
<td>110.87</td>
<td>36</td>
<td>538.23</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>TIME Slope</td>
<td>( r_1 )</td>
<td>1.88</td>
<td>3.55</td>
<td>36</td>
<td>303.73</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Level-1 Error</td>
<td>e</td>
<td>3.55</td>
<td>12.59</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Intercept and Slope references that of the frustrational group while INSTR references values for the instructional group within each Intercept and Slope parameter.

Accuracy Outcome

Set Size 12. Set size 12 results showed a significant mean accuracy score for frustrational students at time zero, \( \beta_{00} = 58.26, t(16) = 6.21, p < .01 \), while instructional students had significantly higher starting accuracy scores at time zero, \( \beta_{01} = 34.30, t(16) = 3.11, p < .01 \),
compared to the frustrational group (Table 5). Additionally, students in the frustrational group showed significant gains in accuracy per intervention session, $\beta_{10} = 4.79$, $t(16) = 3.74$, $p < .01$, while instructional students showed gains that were significantly smaller compared to students in the frustrational group, $\beta_{11} = -4.65$, $t(16) = -3.10$, $p < .01$. Variance results indicated that the variance of the intercepts, 295.45, was significantly greater than 0, $\chi^2 (16) = 47.04$, $p < .01$ suggesting that there are true differences in individuals’ intercepts. The variance in the slopes 1.85, was not significantly greater than 0, $\chi^2 (16) = 14.47$, $p > .05$, suggesting that differences in individuals’ slopes was undetected. Variance attributed to error was 223.11.

**Set Size 24.** Set size 24 results showed a significant mean accuracy score for frustrational students at time zero, $\beta_{00} = 32.56$, $t(18) = 3.36$, $p < .01$, while instructional students had significantly higher starting accuracy scores at time zero, $\beta_{01} = 61.14$, $t(18) = 5.66$, $p < .01$, compared to the frustrational group (Table 5). Additionally, students in the frustrational group had significant gains in accuracy per intervention session, $\beta_{10} = 7.34$, $t(18) = 4.25$, $p < .01$, while instructional students showed gains that were minimal, $\beta_{11} = -7.24$, , $t(18) = -3.85$, $p < .01$.

Variance results indicate that the variance of the intercepts, 260.69, is significantly greater than 0, $\chi^2 (18) = 55.28$, $p < .01$, suggesting that there are true differences in individuals’ intercepts. The variance in the slopes, 4.23, is also significantly greater than 0, $\chi^2 (18) = 30.17$, $p < .05$, suggesting that there are true differences in individuals’ slopes. Variance due to error was 145.91.

Scores using the accuracy metric showed slightly different results compared to use of the fluency metric. Maintained was that students in the instructional group at the beginning of the intervention phase obtained significantly higher scores. Average scores for students in the frustrational group had starting values ranging from 58.26 % accuracy for set size 12 to 32.56 %
accurate for set size 24, while average accuracy scores for students in the instructional group ranged from 92.56 % for set size 12 and 93.70 % for set size 24. Granted, both for fluency and accuracy, this finding is entirely expected because students were placed into groups by their initial performance.

Meanwhile, only students in the frustrational group showed considerable gains in accuracy scores throughout the intervention period. For set size 12, students in the frustrational group showed a significant average increase of 4.79 percentage points per intervention session, while students in the instructional group had an average growth of 0.14 percentage points. Similarly, for set size 24, growth rates were 7.34 percentage points per intervention session for the frustrational group, while students in the instructional group had gains of 0.10 percentage points. Plausible reasons for instructional groups limited gains in accuracy throughout the intervention period is explained in the discussion section.
Table 5

Accuracy Multilevel Analysis Results

<table>
<thead>
<tr>
<th>Set Size 12</th>
<th>Fixed Effects</th>
<th>Parameter</th>
<th>Coefficient</th>
<th>SE</th>
<th>t</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>$\beta_{00}$</td>
<td>58.26</td>
<td>9.38</td>
<td>6.21</td>
<td>16</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>INSTR</td>
<td>$\beta_{01}$</td>
<td>34.30</td>
<td>11.03</td>
<td>3.11</td>
<td>16</td>
<td>.007</td>
<td></td>
</tr>
<tr>
<td>Slope</td>
<td>$\beta_{10}$</td>
<td>4.79</td>
<td>1.28</td>
<td>3.74</td>
<td>16</td>
<td>.002</td>
<td></td>
</tr>
<tr>
<td>INSTR</td>
<td>$\beta_{11}$</td>
<td>-4.65</td>
<td>1.50</td>
<td>-3.10</td>
<td>16</td>
<td>.007</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variance Components</th>
<th>Parameter</th>
<th>SD</th>
<th>Variance</th>
<th>$\chi^2$</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>$r_0$</td>
<td>17.19</td>
<td>295.45</td>
<td>47.05</td>
<td>16</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>TIME Slope</td>
<td>$r_1$</td>
<td>1.36</td>
<td>1.85</td>
<td>14.47</td>
<td>16</td>
<td>&gt;.500</td>
</tr>
<tr>
<td>Level-1 Error</td>
<td>$e$</td>
<td>14.94</td>
<td>223.11</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Set Size 24</th>
<th>Fixed Effects</th>
<th>Parameter</th>
<th>Coefficient</th>
<th>SE</th>
<th>t</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>$\beta_{00}$</td>
<td>32.56</td>
<td>9.71</td>
<td>3.36</td>
<td>18</td>
<td>.004</td>
<td></td>
</tr>
<tr>
<td>INSTR</td>
<td>$\beta_{01}$</td>
<td>61.14</td>
<td>10.80</td>
<td>5.66</td>
<td>18</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>Slope</td>
<td>$\beta_{10}$</td>
<td>7.34</td>
<td>1.73</td>
<td>4.25</td>
<td>18</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>INSTR</td>
<td>$\beta_{11}$</td>
<td>-7.24</td>
<td>1.88</td>
<td>-3.85</td>
<td>18</td>
<td>.001</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variance Components</th>
<th>Parameter</th>
<th>SD</th>
<th>Variance</th>
<th>$\chi^2$</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>$r_0$</td>
<td>16.15</td>
<td>260.69</td>
<td>55.28</td>
<td>18</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>TIME Slope</td>
<td>$r_1$</td>
<td>2.06</td>
<td>4.23</td>
<td>30.07</td>
<td>18</td>
<td>.037</td>
</tr>
<tr>
<td>Level-1 Error</td>
<td>$e$</td>
<td>12.08</td>
<td>145.91</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Intercept and Slope references that of the frustrational group while INSTR references values for the instructional group within each Intercept and Slope parameter.

Reliability

Reliability estimates for the Level 1 intercepts and slopes were also calculated.

Reliability is the ratio of “true” parameter variance to the total observed variance, calculated separately for each individual and averaged over all individuals. High reliabilities indicate a high level of precision in the estimates. For the accuracy predictor, the reliabilities of slope and intercept was low across set size 24, slope = .43, intercept = .69, and set size 12, slope = .22; =
intercept .66. Meanwhile, for the DCPM predictor, the reliabilities of slope and intercept was high across set size 24, slope = .89; intercept = .92, and set size 12, slope = .90; intercept = .87.

Metric reliability is an important component to gauge when determining its use in monitoring performance over time. Metrics that produce values that are unreliable suggest differences in measured score can be attributed to unexplained variables (e.g., measurement error), rather than treatment or student. Reliability coefficients (cronbach’s alpha) greater than .8 are typically considered acceptable. Results indicated that only the fluency predictor produced acceptable reliability coefficients for all students, as reliability was not separately assessed across instructional groups, with reliability results also being consistent across set sizes. Alternatively, the accuracy predictor produced reliability coefficients with limited applicability (i.e., <.80), meaning scores over time were inconsistent and widespread as a likely result of the metric utilized. Of course, differences in reliability coefficients across metrics should not be considered in isolation, but instead assessed with consideration of slope linearity (i.e., slope shape) as further examined in the chapter 5.
Chapter 5: Discussion

Overview

The goal of this study was to investigate an approach to intervention treatment that enhances learning rate efficiency and contributes to overall mathematical intervention research. Due to growing concerns of declining performance in mathematics among grades k-12, schools have rightfully shifted towards implementation of response to intervention (RTI) systems of support to remediate skill deficits through targeted interventions (Gersten et al., 2009). Problematic among traditional RTI frameworks, however, is limited decision-making guidelines that ensures treatment aligns with changing skill proficiency (i.e., instructional match). As students learn new academic skills, their proficiency with the targeted skill(s) also changes and thus so should treatment. Misalignment between skill proficiency and selected treatment potentially hinders overall response rates, as examined in this study. Meanwhile, effects of existing mathematical interventions have not been fully studied with consideration of varying proficiency. Results from the present study therefore contributes to current mathematical intervention research by summarizing a decision-making approach that can be applied to existing interventions to determine whether effectiveness is dependent on proficiency, enhancing overall relevance of instructional match within the context of RTI. In summary, the current study served as a partial replication of prior studies whereby effects of an ET fluency intervention, then supplemented with a feedback component, are examined with consideration of whether initial skill proficiency is aligned with corresponding treatment components (i.e., skill x treatment interaction; Codding et al., 2007; Fontenelle et al., 2020).

A skill x treatment interaction is when intervention components align with initial skill proficiency with the goal of optimizing learning rates (Codding et al., 2007). Evidence of a skill
x treatment interaction would indicate the instructional utility of identifying initial skill proficiency and matching it to an appropriately selected intervention, with secondary gains of specifying whether existing treatments are better suited for students of varying proficiency. In this study, evidence of a skill x treatment interaction would be present if students in the modified instructional group (accurate and/or slow), as this group also included a portion of students from a mastery level (accurate and/or fast), showed greater growth in response to an ET + Feedback intervention than students in the frustrational group (inaccurate and/or more slow). This hypothesis was developed based on results from prior ET studies which showed that ET is most effective for individuals who have already established accuracy (Codding et al., 2007; Fontenelle et al., 2020; Rhymer et al., 2002). The preceding theory then led to the first research question of whether initial instructional level predicted differential treatment response to an ET fluency intervention that was supplemented with performance feedback.

A secondary purpose of this study was to investigate whether observed differences in the treatment response of instructional groups (frustrational/instructional) varied across metrics of accuracy and fluency. Debate continues to exist as to whether instructional level should be measured, and then treatment response monitored, using an accuracy criterion and/or fluency criterion (Burns et al., 2010; Codding et al., 2011). To investigate which metric, accuracy or fluency, better predicts differing response to treatment, instructional groups using both measures was assessed and compared in the context of a skill x treatment interaction. Psychometric properties of each metric, such as reliability, were also compared to assess changes in outcome over time as strong psychometric properties is essential for accurate educational decision-making. High reliability is indicative that outcome variability is likely attributed to treatment, rather than unexplained variables such as tool insensitivity (Burns et al., 2006).
Furthermore, given data reviewed in this study utilized archival data (see Payne et al., 2021), where ET data was collected using probes of different set size, results were examined separately across set sizes to maintain interpretation consistency. Although, trends in response rates across set sizes were compared, outcome measures were separately examined across set sizes because set size is an assessment property that affects probe difficulty, probe variance, and response rates. Recommendations that intervention effectiveness not be evaluated across probes of different set sizes is further supported in research (see Solomon et al., 2019).

**Treatment Response**

To answer the first research question, results indicated that instructional level in some instances predicted differential response to an ET + Feedback intervention, as groups displayed positive treatment growth that was significantly different based on initial proficiency level. However, trends did not fully align with the hypothesis that the instructional group would respond best, relative to the frustrational group, as results were mixed depending on metric use and set size.

**Fluency Metric.** With the fluency metric, the instructional group responded greater to treatment compared to the frustrational group, as predicted, although only for set size 12. For set size 24, the frustrational group responded more, which was not expected. The frustrational group having gains that were greater than the instructional group with respect to set size 24 did not align with previous studies which found that students at an instructional level, relative to those at a frustrational level, responded best to ET (Codding et al., 2007; Fontenelle et al., 2020). Therefore, depending on examined set size, as well as with consideration of sample size adequacy, results partially aligned with previous findings that students in the instructional range
benefit more from an ET intervention than students in the frustrational range (Codding et al., 2007; Fontenelle et al., 2020).

However, results did indicate that both groups positively gained from the ET + Feedback intervention and which was not dependent on set size. This is consistent with previous research findings that have shown that implementation of an ET intervention can produce positive fluency gains across instructional groups, although with rates differing based on instructional level grouping (Codding et al., 2007; Fontenelle et al., 2020). Frustrational learners having benefited from the intervention can be attributed to what Codding et al. (2007) summarized as the positive effects repeated practice has on fluency gains for students of varying proficiency. Even though ET is considered a fluency intervention, and thus considered most appropriate for instructional level learners, this study suggested that individuals in the frustrational level can still benefit with fluency gains. Although other treatments for frustrational learners should first be considered prior to fluency interventions, as it has been shown that frustrational learners have responded better to simple repeated assessments than with ET instructions (Codding et al., 2007).

**Accuracy Metric.** Regarding accuracy, differences in groups response to treatment was more noticeable. For this metric, the frustrational group showed greatest gains to treatment with average increases of 4.79 percentage points per session for set size 12, and 7.34 percentage points per session for set size 24, while students in the instructional group showed nearly no growth, with average increases of .14 percentage points per session for set size 12, and .10 percentage points per session for set size 24. Results again indicated significant differences in groups’ response to the intervention; however, did not align with the hypothesis that the instructional group would respond best. In fact, accuracy results cannot be interpreted within the same skill x treatment hypothesis framework used for fluency as, at the time of this study, no
contemporary skill x treatment interaction literature, using an accuracy metric and an applied experimental design, existed. That is, no previous studies have examined instructional group differences using an accuracy metric, on the basis that accuracy, relative to fluency (DCPM), is not as reliable of a metric to monitor progress of mathematical CBM data (Burns et al., 2006). Observed outcomes are therefore presented to demonstrate reasons why accuracy data might not then be supported as an effective means to ensure instructional match and lend support to fluency as a more valid predictor.

**Metric Utility**

Using accuracy data alone, an educator might assume that ET was ineffective for the instructional group as growth over time was near non-existent. However, a reason why accuracy gains were minimal for the instructional group, supporting fluency as a more useful predictor than accuracy alone, is because an accuracy metric does not capture changes in proficiency for individuals who are already accurate at the start of the intervention. Only with a timing function can two individuals who are both highly accurate be differentiated by proficiency (Binder, 1996). When students in the instructional group are highly accurate to begin the intervention, then early into treatment they reach a ceiling effect as there is limited room for accuracy scores to improve. Thus, students who are highly accurate at the start can only increase proficiency by increasing speed, which is not accounted for using an accuracy metric (Binder, 1996). VanDerHeyden and Codding (2020) displayed the two metrics’ correlation visually by plotting accuracy and fluency scores per individual included in a class-wide math intervention. The authors concluded that DCPM, and thus use of a timing function, was the only way to differentiate individuals who were highly fluent (measured with DCPM) and accurate, from students who were highly accurate but not necessarily fluent. Their analysis further suggested that as individuals’ performance became
more fluent, errors decreased, raising question about the utility of an accuracy metric for fluent learners altogether.

Reliability was another important variable assessed in this study to help determine which metric was the more useful predictor. Results showed that reliability values were higher for the fluency metric compared to the accuracy metric, for both slope and intercept, with results remaining consistent across set sizes. Furthermore, variance of intercept, slope, and variance attributed to error were all higher for accuracy. Fluency reliability being higher than accuracy aligns with previous studies comparing reliability, which has found fluency to be a more reliable metric (Burns et al., 2006). This makes sense when it is considered that accuracy values are determined by the percentage of items correct per total problems attempted. Therefore, variability in performance for frustrational learners would be especially high, resulting in poor reliability, as accuracy scores would fluctuate because individuals are being assessed, and then penalized, on a skill that they have not yet obtained proficiency with. Of course, differences in reliability coefficients should not be interpreted without consideration of slope linearity (i.e., slope shape) as a misfit between the data and slope model can skew reliability coefficients, including their interpretation.

An appropriate growth model (i.e., linear, quadratic, cubic, etc.) is needed when assessing reliability because interpretation of growth parameters may be compromised if a linear model is used to estimate growth that is non-linear (Nese et al., 2011). Examination of accuracy data for the frustrational group indicated near linear growth. However, growth for the instructional group was visually non-linear with high initial accuracy scores (>90%) and minimal growth over time as students preformed at the ceiling. When accuracy data from both instructional groups was then aggregated to examine the reliability coefficient output, overall slope took on a non-linear
shape. That is, growth in accuracy predictably deaccelerated as it approached the ceiling. The misfit between the linear HLM model and the data’s curvilinear shape should be considered as a contributor to why reliability values are greater for the fluency metric as the ceiling is raised with the introduction of a time component. Nonetheless, fluency yielded superiority both in terms of reliability, interpretation, and sensitivity to growth. Of course, interpretation of presented accuracy and fluency reliabilities should be considered in the context of this CBM study, rather than generalize to metric use in other forms of educational decision-making.

An additional variable to consider when interpreting reliability is how it changes in response to the number of data points collected, as well as with changes in instructional proficiency. CBM are considered reliable measures alone but have demonstrated advances in reliability with the collection of multiple data points as it reduces error estimates (Christ & Silberglitt, 2007). The rule of thumb in practice has tended to be that 6–8 weeks of data are sufficient to establish the effectiveness of an intervention (Ardoin et al., 2013). Although mathematical CBM materials, opposed to literacy materials, have emerged to require less time to determine effectiveness attributed to elevated precision (Solomon et al., 2019). With consideration that reliability values increase with the collection of additional data, is should be considered whether the outlined four-week intervention was sufficient to produce accurate reliability estimates. Additionally, when interpreting overall reliability for the fluency metric, it should be considered that reliability is likely to be greater for students performing with higher baseline proficiency as performance of these individuals is to remain more consistent given the skill has been sufficiently obtained. On the other hand, individuals in the frustrational range would be expected to have scores that vary as they learn the skill, but then level out as proficiency improves and errors are reduced. Ultimately, the following changes in performance
across learners and time should at least be considered in the context of reliability estimates, as they will alter reliability values over the course of treatment.

**Baseline Proficiency**

Although not a direct aim of this study, an additional finding apparent was differences in the groups’ initial proficiency at the start of the intervention (i.e., intercept). Not surprising was that students in the instructional group had higher average intercept scores at the start of the intervention than students in the frustrational group. This was because, by definition, students were assigned to either group based on either high or low initial measurement values. However, comparing group intercepts to the cutoffs of Deno and Mirkin’s (1977) fluency criteria and Gickling’s (1977) accuracy criteria, it appeared that the frustrational group had average values that were nearer the low-end of either criterion’s range. For example, using the fluency metric, the frustrational group had average intercepts that were not significantly higher than zero and thus at the lower-end of Deno and Mirkin’s (1977) adapted frustrational range of 0 – 9 DCPM. Whereas the instructional group had average intercept values of 15.74 DCPM for set size 12 and 20.18 DCPM for set size 24, both which were well above the lower-end of Deno and Mirkin’s (1977) adapted instructional level criteria of 10+ DCPM. A similar trend was noticeable for accuracy outputs when compared to Gickling’s (1977) adapted accuracy criteria of < 70 % for frustrational learners and 70 – 100% for instructional learners. Average intercepts values for the frustrational group were 58.26% for set size 12 and 32.56% for set size 24, both which were near the mid- to lower-end of the described accuracy criteria. While average intercepts for the instructional group were 92.56% for set size 12 and 93.70% for set size 24, both near the upper-end of the adapted instructional criteria. Differences in average intercept values across
instructional groups suggest there was a larger distribution of students with varying proficiency in the frustrational range with some likely having varying learning rates compared to others.

A likely reason for differences in the distribution of intercept values from criterion cutoffs was because, specific to this study, students who would have typically measured in the mastery range (20+ DCPM) of the instructional delivery model were recategorized to the instructional range. Therefore, when instructional group average scores were combined with students who scored in what would have been the mastery range, average scores were elevated.

In fact, this study’s sample indicated that for fluency data, 18% of all participants in set size 12 (7 individuals in sample of 38) and 30% of all participants for set size 24 (12 individuals in sample of 39) measured in what would have been initial mastery range. The observation that the inclusion of mastery range individuals elevated the instructional group’s average intercept value was further evident by the reported average intercept of 20+ DCPM for set size 24, which would have normally measured in the mastery range. Given response to ET for students in the mastery range has not been fully studied, it is possible that this group may not have responded well to treatment or may have encountered a ceiling effect soon into treatment, which might then explain why the frustrational group responded best to treatment for set size 24.

**Examining Mixed Results**

Apparent in this study were mixed results, whereby for set size 24, results did not follow the expected outcome that the instructional group would respond best to treatment. This finding could be attributed to collective changes in group response rates across set size conditions. Specifically, the instructional group in set size 24 (3.29 DCPM increase per session) performed worse than the instructional group in set size 12 (4.46 DCPM increase), while the frustrational group in set size 24 (4.64 DCPM increase) performed better than the frustrational group in set
size 12 (2.84 DCPM increase). The instructional group having declines and the frustrational group having gains, both with respect to changes from set size 12 to set size 24, collectively might explain why the frustrational group presented with greater growth compared to the instructional group for set size 24.

Differences in response rates of the instructional group from set size 12 to set size 24 might best be explained by a ceiling effect. The instructional group for set size 24 began the intervention with higher average initial fluency values, in the mastery range. Therefore, soon into intervention this group could have had minimal room for gains as they were already reaching near maximum performance.

The difference in response rates across set size for the frustrational group is a bit more challenging to explain as results counter expected effects of set size on fluency gains. A smaller set size presents a narrower curricular scope (attributed to reduced unique problems repeated) which is associated with high learning rates (Poncy et al., 2015; Solomon et al., 2019; Solomon et al., 2020). However, in the current study, the frustrational group responded best within the larger set size condition. A possible explanation for this outcome is because, although a smaller set size is associated with increased learning rates, the association between learning rates and set size has not been thoroughly examined experimentally with probes derived using a random assortment of problems that participants are expected to know, however, may not. In other words, set size has not been thoroughly examined when content is misaligned with presented skill set, as in the case when a fluency intervention is provided to frustrational learners. In this instance, probe content, such as whether known problems are presented, could likely have greater relevance to overall performance, than what would be attributed to reduced problem set. A larger set size presents more variety of problems to selectively answer, whereas with a smaller
problem set, if the student does not know how to answer the problems correctly, problem repetition would be irrelevant, or potentially even hinder overall performance. Solomon et al. (2019) examined the effects of set size on score precision (i.e., reliability) which indicated that probes of smaller set size had the largest growth rate, but the worst precision, when compared to larger set sizes. That is, although growth rates were largest for the smaller set size, due to increased repetition of unique problems, performance varied because each probe sampled a distinct set of target problems that individuals may or may not have been proficient with. However, this variability eventually leveled out with increased treatment sessions. In summary, set size should be a consideration when determining whether the goal is to maximize response rates or sample a broad degree of target problems, especially when treatment length is limited.

Increased performance of the frustrational group in set size 24, in comparison to set size 12, could then perhaps be explained by problem selectivity that was exacerbated with providing frustrational learners a fluency intervention that is less suitable relative to an acquisition intervention that specifically addressed their needs. When a frustrational learner is provided a fluency intervention, which is not optimized for skill acquisition, performance is largely dependent on the probability that the learner is presented with known items on that day’s specified probe. Original CBM procedures require random sampling of skills from the student’s curriculum (Fuchs and Deno 1991), but a weakness in this approach is variable difficulty of these locally created measures (Fuchs and Deno 1994). Individuals then presented with 24 unique problems (of the possible 81 single digit multiplication problems that omit 0’s and 10’s), as opposed to 12 unique problems, have increased chances of encountering known problems, and then be able to become faster on them, especially if they are simpler single digit-repeating items (e.g., 5x5), which are also likely to be retained across sessions. Specific problems that
frustrational learners then select to answer and which they are instructional with, can result in increased accuracy and speed across sessions, accounting for higher growth. Even if larger set size probes are considered more difficult, performance may be countered with increased probability that more known problems are presented. Furthermore, because DCPM does not penalize for incorrect or skipped problems, although students are encouraged not to skip and were corrected during administration, students completing a probe of a greater set size have more variety of problems to select from. Whether this variability helped or hindered the individual’s performance depended on the individual’s proficiency with the assortment of problems presented. Furthermore, it is not uncommon for individuals to be in tune with their proficiency on a task and be selective in their response processes to enhance motivation as tasks that are too hard may frustrate the learner (Linnenbrink & Pintrich, 2002). This is especially relevant when it is considered that summative error correction was provided per each intervention session, making individuals aware of changes in their performance across sessions.

It is important to clarify that differences in the frustrational group’s performance across set sizes is hypothesized to be attributed to increased problem selection, as a function of set size and the individual’s proficiency with select problems, rather than variability in the difficulty of probe content. This is because it should be assumed that all conditions involved random selection of problems of equivalent difficulty. Thus, assuming items were not intentionally skipped, there should be an equal proportion of easy and difficult problems for both set size 12 and set size 24 conditions. However, the larger set size will inevitably present with more items to select from (both easy and difficult), in which if problems are skipped and familiar items are repeatedly answered and retained across sessions, growth will increase. Furthermore, this hypothesis seems plausible given data on specific problems attempted were not collected as part of this archival
study potentially making problem selection a salient contributor to overall performance, especially for frustrational learners with varying proficiency on select problems.

**Implications**

There are several implications as a result of the findings in this study. One implication is that identifying student’s instructional level prior to intervention application can be a useful process to differentiate treatment effectiveness, especially when used with a fluency metric. Significant differences in response rates of instructional groups were evident, although conditional on set size. Utilization of a skill x treatment interaction demonstrated that, over the course of the 7-assessment period treatment, frustrational learners had total average gains of 19.88 DCPM (for set size 12) and 32.48 DCPM (for set size 24), whereas instructional learners had total average gains of 31.22 DCPM (for set size 12) and 23.03 DCPM (for set size 24).

These findings partially align with previous findings which have shown that smaller set sizes are associated with greater learning rates (Solomon et al., 2020), although such studies continue to lack experimental control over whether treatment content is appropriately matched to presented skill proficiency. Results from this study, as well as that of prior skill x treatment interaction research, raises some evidence for teachers use of the previously outlined decision-making guidelines to ensure that instructional content matches student proficiency with possible implications of increased learning rates.

Another implication of this study is to provide supporting evidence as to why a fluency metric might be superior to accuracy when utilizing CBM data for instructional decision making. The following study demonstrated, through an applied intervention, how a fluency metric is more sensitive to gains in proficiency, reducing ceiling effects observed with an accuracy metric that otherwise would prevent growth from being observed. With a metric that is more sensitive to
gains in proficiency, teachers can more accurately rely on the data to inform treatment response and thus educational decision-making. Decision making accuracy is also enhanced with increased reliability, which was found with fluency data, indicating that changes in student growth is likely attributed to treatment variables, rather than some unexplained variance, such as variability in probe content or how scores are calculated. Furthermore, fluency can serve more useful in assisting teachers to differentiate proficiency among students who are accurate and slow vs accurate and fast. Students who are accurate but very slow would be considered in the instructional range when an accuracy metric is used. However, using a fluency metric, and depending how slow response time is, the same student may test in the frustrational range. With the accuracy metric, this student would be considered to have adequately learned the skill and is ready to work on gaining speed, making a fluency intervention appropriate. However, using a fluency metric, this student would be considered acquiring the skill and would benefit from an acquisitional intervention as the student is likely using a strategy to solve the problem that does not rely on automatic recall.

Although results from this study suggest fluency as the optimal predictor in the context of CBM, it does not necessarily mean that accuracy should be discarded entirely as a scoring metric within education. Determining the appropriate use of either a fluency or accuracy metric partly depends on the stage of proficiency assessed, as well as the purpose of assessment procedures. If individuals assessed reside in the frustrational stage, then an accuracy metric would be considered appropriate as the goal at this stage is to build accuracy, will less of an emphasis on speed. Furthermore, if the aim of assessment is to gauge what students are expected to have already learned (i.e., mastered), an accuracy metric would be suitable. This form of assessment is summative with the goal being to evaluate learning outcomes that students are expected to
already be proficiency with (Dixson & Worrell, 2016). However, if assessment is more intermediate in frequency and used to gauge the learning process, information that could then be used as feedback to inform teaching, it would be optimal to utilize a fluency metric as it is more sensitive to changes in proficiency, and thus the learning process. This approach to formative assessment aligns with the intents of CBM, which is to monitor changes in proficiency throughout the intervention process and make changes to treatment accordingly (Dixson & Worrell, 2016). The following conclusions imply that the instructional utility of metrics, along with measurement interpretation, can change over time and context depending on inferences in which proficiency has been met. When instruction has been deemed sufficient to produce proficiency that is comparable to outlined benchmarks, then the interpretation of metric results can change, such as the interpretation of accuracy values when used in the context of either formative assessment or summative assessment.

**Limitations and Future Research**

One limitation to this study was that mastery level performance was recategorized into the instructional group. While comparing outcome measures of an ET + Feedback intervention for frustrational and instructional learners was the primary purpose of this paper, the integration of mastery level participants should not be overlooked as there is limited evidence of the effectiveness of ET for this population. Similar to that observed with the accuracy metric for the instructional group, students in the mastery range for the fluency metric may show differential response to treatment as they arrive at the measurement’s absolute or functional ceiling. Meanwhile, according to the instructional hierarchy, individuals who demonstrate skill mastery do not require a fluency intervention altogether but instead should receive an intervention that promotes skill generalization - the next stage in skill development (Haring et al., 1978).
A future study examining the effects of a skill x treatment interaction in the context of an ET intervention might want to then examine all three instructional groups (Frustrational, Instructional, Mastery) response to treatment, as opposed to just two. This would support each group’s independent response to treatment, removing the confounding variable that group differences in response rates are attributed to aggregated members with varying proficiency. This is especially relevant given the limited research of mastery range students response to ET.

Due to this paper being archival data, only a portion of probes scored for fluency (n = 77) were also scored for accuracy (n = 38). Additionally, while sample size was comparable across set size for each metric, students were not evenly distributed across instructional groups. Due to smaller sample sizes across instructional groups, group averages were more sensitive to changes in individual scores, which may have affected overall response rates. Reduced power, due to smaller sample sizes, as well as mixed results that vary from the limited studies that have examined the effects of ET across instructional levels, again supports additional need for replication research.

As statistical power was not calculated as part of this study, low sample size and unknown power, can limit the confidence in which conclusions are drawn from presented results. Statistical power is the probability of detecting an effect when there is a true effect present to detect (Shadish et al., 2002). Power is increased with greater sample size and supports confidence in the conclusions drawn from the results as it reflects the probability of finding an effect if there is an effect to be found. Experimental results with too low statistical power will lead to invalid conclusions about the meaning of the results. Because power was not explicitly measured in this study, raising concern as to whether sample size was sufficient to establish confidence in outcomes, sample size should be considered a possible limitation to the
significance of the conclusions drawn. Although, it should be noted that the sample size for this study was comparable to prior to studies that have examined a skill x treatment interaction, such as that by Codding et al. (2007), which included 98 participants distributed across three distinct intervention conditions.

This study being archival posed another limitation as there was reduced experimental control over described variables. While archival data can be helpful, especially in the context of meta-analyses, it also has limitations. Among these limitations are missing pertinent data, data that is fixed within the parameters of prespecified conditions, and difficulties of establishing causal inferences, among others. In this case, archival data used as part of this study did not include the probes completed per assessment session. Information from these probes could have served useful for item analysis and in determining the degree to which non-attempted, incorrect, and repeated responses affected overall fluency scores, which would have lent insight to whether item sampling and problem selection had relevance to varying performance across set size conditions (see Solomon et al., 2019).

A fully experimental study would correct for the limitation of this study being archival. Two important elements that would have been helpful to include in this experimental study is a control group and an additional intervention treatment that is suitable for frustrational learners (e.g., CCC). A control group would serve as a benchmark to compare to the treatment group to better understand the effect of intervention effects. In this case, the control group might entail a portion of students who receive no intervention, but instead participate in repeated assessments, with scores then being examined across instructional levels and treatments. Inclusion of an additional intervention suitable for frustrational learners, such as CCC, in addition to a control group, would make this a 3x2 experimental study and increase validity. In this case, frustrational
learners would be matched and mismatched to an acquisition intervention and instructional
students would be matched and mismatched to a fluency intervention, which results then being
compared to outcomes of the control group. If frustrational learners were to better respond to the
acquisitions intervention, and instructional learners better respond to the fluency intervention,
there would be increased evidence of how skill x treatment procedures support differing learning
rates when specified interventions and its components are aligned with initial skill proficiency
levels. Furthermore, such experimental design would want to include random assignment of
members who are equally distributed across the proposed conditions, correcting for differences
in sample size.

Coddin et al. (2007) included both a control group and a CCC acquisition intervention
to compare against ET effects. The authors found that CCC, ET, and a control treatment
produced increases in fluency as all conditions promoted increased opportunities to respond.
Treatments effects were greatest for instructional learners who received ET, while effects were
least optimal for frustrational learners who received ET. That is, frustrational learners in the
control group responded better than frustrational learners in the ET group. The authors attributed
these differing effects to a timing component. Participants were only explicitly aware of the
timing component in the ET treatment as it was part of intervention instructions. Time was
assumed to have increased motivation for instructional learners by motivating them to reach
mastery, however, reduced the effects of treatment gains for frustrational learners. Meanwhile,
while it was expected that frustrational learners would have responded best to CCC, these
individuals responded no better than the control group. Differences in instructional groups
varying response to treatment, especially the frustrational group who received a well-studied and
effective acquisition intervention (Skinner et al., 1989; Poncy et al., 2007; Poncy et al., 2012),
again highlights relevance of additional replication in the field of skill x treatment interaction research.

Due to mixed results from this study and limited research investigating the effects of a skill x treatment interaction, additional replication is warranted. However, exact replication of this study may not be feasible given the array of experimental conditions that were partly in place due to original study design. With consideration of the direction for future research, it is worth highlighting research questions that remain to be investigated as result of this study’s outcomes. One question that remains is whether students in the mastery range have differential response to ET. Determining treatment response for this group is relevant in determining whether students with this level of proficiency would even benefit from continual exposure to a fluency intervention or should instead move on to a more advanced skill entirely, as recommended by the instructional hierarchy. Another question that remains is determining the optimal set size that supports sufficient treatment rates and reliability, but also samples an adequate selection of problems that the individual is expected to become proficient with. While reduced set size has demonstrated to increase learning rates, it has also been shown to reduce precision, which reflects a tradeoff between maximizing response rates and sampling a sufficient degree of problems within the selected skill. Therefore, research to determine an appropriate set size, that both demonstrates sensitivity to growth, but also exposes individuals to a problem set with breadth, is warranted. Furthermore, it might be helpful to investigate the reliability of accuracy purely for frustrational learners. Provided the goal of frustrational learners is to build accuracy, an accuracy metric would be sufficient to monitor progress for this group so as long as the metric is reliable.
Chapter Summary

Evidence of a skill x treatment interaction supports consideration of aligning student abilities with specified intervention components to strengthen overall treatment effects. As partially evident in this study, when skill proficiency is appropriately matched to treatment, response rates can be optimized, lending benefit to enhancing the effectiveness of skill remediation supports.

Results from this study also lends benefits to the degree in which fluency is a more useful predictor compared to accuracy. In an applied experimental design, this study displayed how fluency better differentiates students by proficiency, produces a more reliable measure of outcome, and is sensitive to changes in proficiency in response to treatment, all of which enhances instructional decision-making. Because of its timing function, fluency is the only predictor that can differentiate two students who are both highly accurate but respond at different rates. These differences in rate are useful in determining differences in proficiency, such as differentiating the student who has obtained accuracy, but is slow, and thus entering the instructional level, from the student who has obtained accuracy, but is fast, and thus entering the mastery level. Fluency also provides greater reliability (as shown in this study), in comparison to accuracy, meaning that scores per session reflect a more accurate representation of growth, with this information then serving as the basis for treatment continuation or change. Fluency, due to its raised ceiling which can be attributed to its timing function, it also the only metric that is sensitive to progress for when maximal accuracy has been reached. In this study, if accuracy was to be solely used to determine treatment response, the instructional group would have demonstrated no effects to treatment. However, results in this study demonstrated this group did benefit with proficiency gains, but with respect to completing more problems and at faster rates.
To support increased understanding of skill x treatment processes (e.g., instructional hierarchy), additional replication is needed, especially among understudied content areas like mathematics. This includes how initial proficiency affects response to varying treatments, in addition to how metric use affects decision making, as both are pertinent to the integrity and relevance of skill x treatment processes. As instructional match is increasingly studied within applied studies, the effects of previous interventions can be reexamined to determine whether they are more optimal to groups of differing instructional level. For example, Fontenelle et al (2020) examined whether instructional groups responded differently to a Taped Problems intervention, which was also compared to ET. In partial testament for the need to further study skill x treatment interactions, the current study showed that students of varying proficiency present with varying response to ET treatment. Although, results were mixed as likely attributed to other understudied factors such as set size. This study also demonstrated, in an applied case, increased reliability of using a fluency metric to interpret treatment response, which can better guide future educational decision making. In summary, the results suggested that initial instructional level can serve useful in predicting response to an ET + Feedback intervention, however, because results varied across set size conditions, additional replication is warranted.
References


https://doi.org/10.3102/00028312021002449


https://www.hmhco.com/programs/go-math#overview


### Appendix A: Administration Probe Exemplar (Set Size 12)

Name: ____________________

#### 1 X 1 Multiplication

<table>
<thead>
<tr>
<th></th>
<th>7</th>
<th>4</th>
<th>4</th>
<th>10</th>
<th>7</th>
<th>2</th>
<th>5</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>8</td>
<td>10</td>
<td>9</td>
<td>9</td>
<td>5</td>
<td>3</td>
<td>6</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>7</th>
<th>8</th>
<th>6</th>
<th>2</th>
<th>2</th>
<th>7</th>
<th>6</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>8</td>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>8</th>
<th>4</th>
<th>2</th>
<th>7</th>
<th>7</th>
<th>3</th>
<th>10</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>6</td>
<td>9</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>10</th>
<th>6</th>
<th>7</th>
<th>2</th>
<th>4</th>
<th>4</th>
<th>7</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>9</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>9</td>
<td>10</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>5</th>
<th>7</th>
<th>8</th>
<th>2</th>
<th>4</th>
<th>5</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>3</td>
<td>10</td>
<td>6</td>
<td>8</td>
<td>9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>6</th>
<th>2</th>
<th>10</th>
<th>7</th>
<th>7</th>
<th>2</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>2</td>
<td>4</td>
<td>9</td>
<td>8</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>7</th>
<th>4</th>
<th>4</th>
<th>7</th>
<th>7</th>
<th>10</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>4</td>
<td>2</td>
<td>9</td>
<td>10</td>
<td>2</td>
<td>8</td>
<td>9</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>6</th>
<th>10</th>
<th>7</th>
<th>6</th>
<th>10</th>
<th>6</th>
<th>3</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>2</td>
<td>9</td>
<td>8</td>
<td>2</td>
<td>9</td>
<td>2</td>
<td>8</td>
<td>2</td>
</tr>
</tbody>
</table>
Appendix B: Administration Probe Exemplar (Set Size 24)

Name: ______________________

1 X 1 MULTIPLICATION

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>6</td>
<td>7</td>
<td>5</td>
<td>6</td>
<td>9</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>x</td>
<td>5</td>
<td>x</td>
<td>3</td>
<td>x</td>
<td>10</td>
<td>x</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>8</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>x</td>
<td>5</td>
<td>x</td>
<td>3</td>
<td>x</td>
<td>3</td>
<td>x</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>9</td>
<td>10</td>
<td>7</td>
<td>9</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>x</td>
<td>8</td>
<td>x</td>
<td>9</td>
<td>x</td>
<td>2</td>
<td>x</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>10</td>
<td>7</td>
<td>10</td>
<td>2</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>x</td>
<td>9</td>
<td>x</td>
<td>3</td>
<td>x</td>
<td>5</td>
<td>x</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>9</td>
<td>8</td>
<td>6</td>
<td>10</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>x</td>
<td>6</td>
<td>x</td>
<td>3</td>
<td>x</td>
<td>3</td>
<td>x</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>9</td>
<td>3</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>x</td>
<td>7</td>
<td>x</td>
<td>6</td>
<td>x</td>
<td>8</td>
<td>x</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>7</td>
<td>2</td>
<td>10</td>
<td>6</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>x</td>
<td>3</td>
<td>x</td>
<td>3</td>
<td>x</td>
<td>10</td>
<td>x</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>9</td>
<td>6</td>
<td>10</td>
<td>5</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>x</td>
<td>3</td>
<td>x</td>
<td>3</td>
<td>x</td>
<td>2</td>
<td>x</td>
<td>9</td>
</tr>
</tbody>
</table>
Appendix C: Explicit Timing Administration Protocol

Administration Protocol – Explicit Timing

Treatment Integrity Checklist

- Give treatment integrity checklist to teacher and ask the teacher to complete it.

Assessment

- Pass out the folders to students. Make sure each student gets the folder with his/her name. Say, “Please keep your folder closed until I tell you to open it.”

- Read the following directions: “Open up your folder. Take out your math worksheet from the left side pocket. Close your folder.” Scan the room to make sure everyone has their math sheet out and the folder closed. “This is your math sheet with multiplication problems. Some of the problems may be easy and some may be difficult. Start with the first problem and work across the page. Go in order and don’t skip any. If you come to a problem you do not know, put an X through it after you have tried it and go to the next problem. Keep your eyes on your math sheet and do not use any calculators or other aids. You will have 1 minute to complete as many problems as you can. Keep working until I tell you to stop. Ready? Begin.” Begin timing.

- Walk around the room to make sure students are completing the page correctly (i.e., going in order, not skipping problems). Redirect students as needed. Praise students who are on-task.

- After 1 minute, say, “Stop. Put your pencils down.” Reset stopwatch.

- Look around the room to make sure students have stopped working. Say, “Please put your math sheet back in the left pocket of your folder.”

Intervention

- Say, “Now take out your math sheet from last time from the left side pocket. Any problems you got wrong are marked in red. Look at each problem marked in red and whisper the problem and correct answer softly to yourself. You have 1 minute. Go.” Begin timing 1 minute.

- “Put your math sheet from last time back in the left pocket of your folder. Take out the new math sheet for today from the right side pocket. Close your folder. Write your name and today’s date at the top of your math sheet.” Scan room to make sure everyone has their ET sheet out and the folder closed. “Now I am going to give you 4 minutes to complete as many problems as you can. Try to complete each problem correctly and do not skip around. If you finish the first page, turn it over and continue working. Keep working as quickly as possible until I tell you to stop. Ready? Begin.” Begin timing.

- Walk around the room to make sure students are completing the page correctly (i.e., going in order, not skipping problems). Redirect students as needed. Praise students who are on-task.

- After 4 minutes, say, “Stop. Put your pencils down.” Look around the room to make sure students have stopped working. Say, “Please put your math sheet back in your folder.” Collect all folders and the treatment integrity sheet from the teacher.