Comparing Schedules of Progress Monitoring Using Curriculum-Based Measurement in Reading: A Replication Study

By

Samantha A. Gesel

Dissertation

Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in
Special Education

August 9, 2019
Nashville, Tennessee

Approved:

Christopher J. Lemons, Ph.D.
Lynn S. Fuchs, Ph.D.
Elizabeth Talbott, Ph.D.
Paul J. Yoder, Ph.D.

ACKNOWLEDGEMENTS

The research described in this article was supported in part by Grant H325H140001 from the Office of Special Education Programs, U.S. Department of Education. Nothing in the article necessarily reflects the positions or policies of the federal government, and no official endorsement by it should be inferred.

The research described in this article was also supported in part by Peabody College of Vanderbilt University's Professional Development Funds and the Department of Special Education's Melvin I. Semmel Award for Dissertation Research. Without all of these sources of financial support, this work would not have been possible.

Thank you to my research assistants, who truly made this research study go as smoothly as it did. Their dedication to the study and the children with whom they worked was inspiring. Thank you also to Dr. Lynn Fuchs, Dr. Paul Yoder, and Dr. Betsy Talbott. As committee members, their guidance throughout this study was invaluable. Finally, I would especially like to thank my advisor, Dr. Chris Lemons. I am incredibly grateful for all of his mentorship and support throughout my entire doctoral program.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Special educators are tasked with providing specialized and individualized academic services to students with severe and persistent learning difficulties (Individuals with Disabilities Education Improvement Act [IDEIA], 2004). To determine students' intervention needs, special education teachers must collect and analyze student data, assess student responses to existing intervention protocols, and evaluate the need for instructional changes if students demonstrate inadequate growth (Danielson & Rosenquist, 2014; Gersten et al., 2008). Recently rebranded as data-based individualization (DBI; www.intensiveintervention.org), the process of data-based decision making was first described in the work of Deno and colleagues in the 1970s (Deno & Mirkin, 1977).

The DBI Process

The DBI process is iterative and involves using data to intensify and individualize interventions so that students may meet instructional goals. The National Center on Intensive Intervention (NCII) provides a model of the DBI process that includes five steps (see Figure 1). First, teachers select a validated, evidence-based intervention. The intervention should be aligned with a student's academic or behavioral needs, as it serves as a foundation for his or her intensive intervention. Second, teachers regularly collect and analyze students' data to monitor student progress and responses to the evidence-based intervention. Third, for students whose data indicate inadequate response, teachers conduct diagnostic assessments. Teachers use multiple

data sources and analyze student errors to begin to consider the reason an intervention may not be working for a student. Fourth, teachers use the diagnostic data to plan and implement systematic intervention adaptations. These adaptations should intensify the existing, standard intervention protocol, and should be individualized based on the student's needs. Fifth, teachers monitor student progress to determine if a student's response to intervention improves after implementing planned adaptations. For students whose data continue to show inadequate response, teachers continue this iterative, problem-solving approach. Teachers repeat the third through fifth step as needed to address inadequate response and improve academic outcomes.

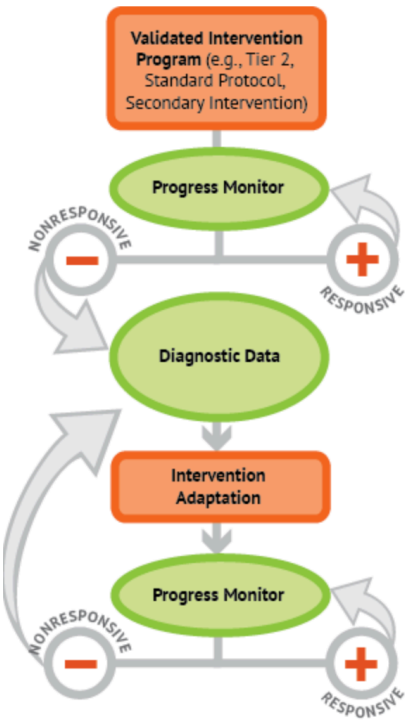

Figure 1. A data-based individualization (DBI) model from the National Center on Intensive Intervention (NCII; www.intensiveintervention.org).

Progress monitoring and curriculum-based measurement. Ongoing data collection through progress monitoring (PM) is essential to DBI and the broader aim of special education (Gersten et al., 2008). Curriculum-based measurement (CBM; Deno, 1985) is a common PM framework used to monitor students' academic progress. CBMs are brief, general outcome measures that assess student performance on an academic skill through the use of multiple, equated assessment probes (Deno, 2003). The use of PM data such as CBM is a pivotal component in the second and fifth step of the DBI model. Student performance on CBM probes provides evidence for the effectiveness of standardized intervention protocols as well as interventions in which teachers have systematically introduced instructional adaptations. To monitor students' reading progress, teachers often use an oral reading fluency CBM, given the strong relation between the number of words read correctly (WRC) and reading achievement (Fuchs, Fuchs, Hosp, & Jenkins, 2001; Reschly, Busch, Betts, Deno, & Long, 2009). CBM vendors publish oral reading fluency CBMs for students reading at a first through eighth grade instructional reading level (e.g., AIMSWeb; Shinn, Shinn, & Langell, n.d.)

Use of CBM to inform instruction: Effect on student outcomes. The research base for DBI and CBM is extensive. It provides a strong evidence-base for the use of these practices in schools to positively impact the academic growth for students with disabilities (Stecker, Fuchs, & Fuchs, 2005). In a narrative review, Stecker et al. (2005) reported that CBM-based interventions have significant positive effects on students' academic achievement in reading, mathematics, and spelling. When teachers use CBM data – displayed in reports with or without additional instructional supports such as instructional recommendations – to engage in the iterative DBI process, they create a powerful framework in which students with the most persistent and challenging needs can demonstrate academic gains.

The results of a more recent meta-analysis updated and supported the evidence that DBI processes work for students with disabilities (Jung, McMaster, Kunkel, Shin, & Stecker, 2018). The authors of this review reported an effect size of g=0.37 for interventions in which teachers individualized instruction based on CBM-reports that provided strictly student performance data (classified by the authors as DBI-only). Additionally, the authors reported an effect size of g = 0.38 for interventions in which teachers individualized instruction based on CBM reports that included additional information, such as instructional recommendations (classified as DBI-Plus). The results of this meta-analysis underscore the importance of collecting, evaluating, and linking CBM data to instruction for students with disabilities.

PM schedules: Balancing accuracy and timeliness. Even with the strong evidence-based supporting teachers' use of CBM and DBI for students with the most intensive needs, teachers must consider the specific PM schedule they plan to implement with students. PM schedules may differ in frequency of data collection (e.g., weekly, biweekly, or more intermittent data collection) and the number of PM probes administered at each data collection timepoint (e.g., one probe per session vs. three probes per session). These factors, in addition to duration of data collection and variability of data, affect the accuracy and precision of the assessment as an estimate of students' true reading achievement (Christ, 2006; Christ, Zopluoglu, Monaghen, & Van Norman, 2013). At the same time, the schedule of data collection teachers employ dictates the amount of data available to analyze. The availability of data affects the frequency with which teachers may engage in data-based decision-making. Consequently, the schedule of data collection affects how frequently teachers may be able to systematically change instruction. This has the potential to affect the timeliness of data-based decisions.

When deciding on data collection plans, researchers and teachers should consider both accuracy *and* timeliness of PM schedules, though striking this balance can be complicated and challenging. Scores on CBM probes provide an estimate of students' reading performance, which approximates students' true reading ability. Examining growth in CBM scores across time provides an estimate of students' true growth in reading-related skills such as reading fluency (i.e., WRC), which approximates students' true growth in their reading ability. Observed CBM scores, however, are limited by the technical adequacy of the assessment and the presence of measurement error above and beyond true score variance. CBM scores vary in the accuracy with which they reflect students' true scores or true growth. Aggregating multiple scores (within and/or across sessions) increases the stability of measures and, therefore, the confidence that those measures reflect true scores (Yoder, Lloyd, & Symons, 2018). Increasing stability of CBMs, however, takes time (Christ, 2006; Christ et al., 2013). For CBM and the DBI process, lower accuracy thresholds take less time to meet, but lead to higher proportions of inaccurate decisions regarding the adequacy of a student's response to intervention (Jenkins, Schulze, Marti, & Harbaugh, 2017). Conversely, higher accuracy thresholds take longer to reach, but provide greater confidence that the data-based decisions reflect students' true growth (Jenkins et al., 2017). Teachers need to select a PM schedule that is sufficiently accurate, while maintaining a level of timeliness that does not interfere with their ability to engage in DBI for students who are inadequately responding.

Current DBI initiatives. Deciding upon a data collection plan and PM schedule is only the first step, however. Stecker et al. (2005) noted that it is not enough to collect CBM data without the systematic use of data to inform practice. Rather, student success is contingent upon

the entire DBI process (i.e., having *a priori* decision rules, reflecting on student progress and errors, and tailoring interventions or goals to the needs of inadequately responding students). The challenge is that teachers often do not collect CBM data or do not use these data to appropriately adapt instruction or instructional goals (Deno, 2014). Even with intensive levels of support to assist teachers in CBM administration, many teachers fail to make any data-based intervention or instructional goal changes (Stecker et al., 2005).

Given that the DBI process is pivotal to the field of special education, this lack of DBI use in schools is concerning. This concern is elevated when framed within the context of the most recent data from the National Assessment of Educational Progress (2017), which indicated that only 12% of fourth grade students with disabilities had met grade level reading benchmarks. The remaining 88% of fourth grade students with disabilities performed at a basic or below basic level of reading. These students may require intensive intervention, such as those provided through the DBI process.

Recent federal initiatives such as NCII (www.intensiveintervention.org) have sought to improve school-based DBI frameworks and support the regular use of DBI practices (Lemons, Sinclair, Gesel, Gandhi, & Danielson, 2019). During its first five-year funding cycle, NCII worked with school professionals in 26 schools by providing technical assistance in the implementation of DBI frameworks. Lemons et al. (2019) reported on lessons learned during these five years based on the content of interviews with school personnel. Overall, the work with NCII led to school- and district-level changes to the frequency with which school professionals engaged in DBI processes. Additionally, the interviewed professionals spoke positively about DBI implementation and the DBI process as a whole. However, school professionals were slow

to integrate DBI implementation into larger school contexts, particularly in a way that would positively improve outcomes for many students across all academic areas.

During their interviews, the school professionals noted challenges in using DBI in their schools (Lemons et al., 2019). Challenges included difficulty in implementing DBI prior to ensuring that the general education programs and standard protocol interventions consisted of evidence-based practices implemented with high quality. Interviewees also acknowledged the need for supportive leadership and strong school-wide dedication to the DBI process. Both of these systems allow for a context in which components of DBI are valued. These contexts include allowing teachers' schedules to include time dedicated to administering and evaluating student data to make instructional adaptations. Finally, school personnel described how the DBI process is challenging and takes a lot of dedication and time prior to seeing student growth. These challenges highlight potential barriers to CBM and DBI implementation and provide insight into why, despite the strong evidence-base for these processes, teachers are not adequately engaging in them.

Considering Alternative PM Schedules to Address Teacher-Reported Time Barrier

In a recently published study, Jenkins et al. (2017) evaluated the relative accuracy of different PM schedules for CBM in reading. Jenkins and colleagues acknowledged that teachers often cite time involved in collecting PM data as one barrier to their use of CBM to inform instructional decisions (Deno, 2003). Jenkins and his team argued that, compared to the traditional, weekly PM schedule, using more intermittent PM schedules may reduce the time it takes to set up data collection contexts and minimize the interruptions PM testing causes to typical instruction. Additionally, the authors argued that decreasing the total number of

measurements would further reduce the time commitments required for CBM testing. In this way, Jenkins et al. framed their investigation of alternative schedules for PM as one way to enhance the feasibility of CBM administration for teachers. Jenkins et al. argued that this would provide more time for teachers to use the data to inform their instruction, particularly for students whose data showed inadequate response to intervention.

Jenkins and colleague's research questions. Jenkins and colleagues had two research questions. First, they considered whether intermittent PM schedules were less accurate than the current, weekly PM schedule standard. Second, they explored the number of weeks it took different PM schedules to reach 70 and 75% accuracy thresholds.

The original sample, data collection procedures, and data analysis. Jenkins et al. recruited 11 special education teachers and 66 students with high incidence disabilities for their study. After excluding students who missed more than one week of data collection, Jenkins et al. had a final sample size of 56 students, 20 of whom were girls. Table 1 provides additional details Jenkins and colleagues reported about their sample of students. Jenkins et al. administered multiple oral reading fluency CBM probes each week of the study. For each student, Jenkins et al. randomized a set of 33 AIMSWeb (Shinn et al., n.d.) and nine Edcheckup (http://www.edcheckup.com) passages. Each student read passages at his or her instructional reading level, as determined by the student's special education teacher. Jenkins and colleagues administered three passages during baseline, three during Weeks 1-11, and six in the final week (Week 12).

Jenkins et al. set a goal growth rate of 1.0 WRC increase each week to determine students' adequate progress, citing this rate of growth as a reasonable standard for students in 2nd through 6th grade that had been suggested by Deno, Fuchs, Marston, and Shin (2001). First, Jenkins and colleagues estimated students' "true growth" by inputting student scores on all 42 CBM probes administered into an ordinary least squares (OLS) regression to calculate a true growth slope. They assessed whether each student's true growth slope indicated adequate (1.0 WRC increase or more per week) or inadequate (less than 1.0 WRC increase per week) progress. Next, they simulated six different PM schedules (i.e., one a week, two Every-2 weeks, and three Every-3, -4, -5, and -6 weeks) by selecting the CBM data points that would have aligned with each PM schedule, had the researchers only collected data according to each PM schedule. Using OLS regression, Jenkins et al. calculated the weekly slope of each PM schedule across the weeks of the study. They used the weekly slopes for each PM schedule to determine whether the schedule's data indicated adequate or inadequate progress, relative to the goal of 1.0 WRC increase. Finally, Jenkins et al. determined the accuracy of each PM schedule by calculating the proportion of students for whom the PM schedule's determination of the adequacy of student growth matched the determination reflected in the true growth data. They compared PM schedules' accuracy across the weeks of the study and reported the number of weeks it took each schedule to reach 70 and 75% accuracy.

Jenkins and colleague's results. Overall, Jenkins et al. reported that PM schedules had similar levels of decision-making accuracy and that more intermittent PM schedules did not undermine timeliness. The authors interpreted their results as demonstrating that the instructional decision-making accuracy and timeliness using data from more intermittent PM schedules was

comparable to and similarly accurate as using data from the traditional, weekly PM schedule. They concluded that this provided evidence for the use of the intermittent PM schedules, and suggested that presenting these intermittent schedules as options to teachers could address time as the primary teacher-reported barrier to engaging in data-based decisions.

Purpose

The purpose of this study was to replicate and extend the work of Jenkins and colleagues, given the impact of this line of research on the field of special education (i.e., wide use of CBM, the potentially controversial results reported in the original study, and the potential impact of the study's conclusions on schools' use of CBM). Replication is an important component of the empirical process, as it adds to the evidence-base for scientific findings and contributes to the understanding of broader theories (Coyne, Cook, & Therrien, 2016). In special education, replication research is drastically underrepresented in the literature base. In a recent review, Lemons, King, Davidson, Berryessa, and Gajjar (2016) reported that replication studies represent only 0.41% of articles published in special education journals.

Direct research questions. I considered the same two research questions as the original study as my primary, direct replication research questions. These included, "Is decision-making accuracy from intermittent PM inferior to that from weekly PM, the current standard?" (Jenkins et al., 2017, p. 45), and "How many weeks of PM do these schedules require to reach specific levels of decision accuracy?" (Jenkins et al., 2017, p. 45). I hypothesized that the decision-making accuracy from intermittent PM would be indeterminately different from that of weekly

PM. I also hypothesized that it would take approximately six weeks for PM schedules to reach 70% accuracy and ten weeks for PM schedules to reach 75% accuracy.

Extension research questions. I also included conceptual replication research questions, which served as extensions to the direct replication. I extended the first research question by considering whether each PM schedule's decisions on the adequacy of student growth was at or above the *a priori* accuracy thresholds of 70% and 75% accuracy compared to students' true growth determination. I only contrasted PM schedules when at least one schedule met the required threshold.

I extended the second research question in two ways. First, I considered whether the time it took intermittent PM schedules to reach each accuracy threshold was within two weeks or less of the time it took weekly PM schedules to reach the same accuracy threshold. I selected the two-week criterion as I hypothesized that instructional changes made at any point within this brief window of time would not lead to differences in student outcomes. I hypothesized that the timeliness of each intermittent PM schedule would be within two weeks of the weekly PM schedule. I also planned to statistically test the difference between PM schedules' time to accuracy thresholds and hypothesized that there would not be a significant difference between intermittent and weekly PM timeliness. This analysis, however, did not end up being possible to conduct (see "Supplementary Data Analyses" subsection of the methods, p. 26).

Replication-related research questions. Finally, I compared the results of this replication study to the original results. First, I assessed whether there was a direct replication of findings based on whether the interpretations of results were the same as those made in the

original study. Second, I assessed whether the interpretations changed when accounting for teacher-reported instructional changes.

Pre-registration. In October 2018, I pre-registered my research plan for this replication (intended sample, data collection procedures, and data analysis) with the Open Science Foundation (OSF). In this pre-registration, I also described the aspects of the original study's methodology that I intended to directly vs. conceptually replicate, planned extensions to the original study, known differences between the planned replication and the original study procedures, and the anticipated effects of those differences. The pre-registration is publicly available at https://osf.io/udxqn. The pre-registration document uses "indirect" replication as the terminology to represent the conceptual replication (or extension) aspects of this study.

CHAPTER 2

METHODS

In the following section, I first describe the study sample, PM materials, procedures, design, and analysis. For ease of comparison across studies, I wrote these descriptions in a parallel format to the way Jenkins and colleagues wrote their methods section. Additionally, I report the key differences across the two studies for transparency in the direct and conceptual replication components of this study.

Sample

I recruited 12 special education teachers from six elementary schools within an urban district in the southeast United States. All teacher participants worked predominantly with students identified with high-incidence disabilities. The majority of the participating teachers were white (*n*=9; 75%) and female (*n*=11; 91.7%). Participating teachers helped recruit 64 students for participation in this study. This sample of participating students is smaller than initially planned in the OSF pre-registration for this study. Because of this, I aim to recruit a second cohort of participants, but will report on the preliminary data from this first cohort for the purpose of this manuscript.

Six of the participating Cohort 1 students moved before the start of data collection. Two additional students moved in the middle of data collection. Finally, five students had poor attendance (i.e., two or more weeks of data collection missed). Following the data cleaning procedures outlined by Jenkins et al., I removed these students from final data analyses. This left

a final sample of 51 students (14 female; 27.45%). Of the 51 students, 42 (82.35%) were present for all 14 weeks of data collection; nine (17.65%) students missed one week of data collection. Results from a *t*-test indicated that the mean difference between true growth slopes for students with incomplete vs. complete data (0.61 and 0.88, respectively) was not statistically significant ($t(49)=-1.38$, $p=0.17$). I calculated the standardized mean difference effect size for these data, using an online effect size calculator (https://www.campbellcollaboration.org/escalc/html/EffectSizeCalculator-SMD2.php). There was an effect size of $d=-0.5069$ [-1.2335, 0.2197] between true growth for students with incomplete vs. complete data. This suggests that, though the difference was not statistically significant, students with missing data, on average, demonstrated poorer true growth than students with complete data. This poor growth could be due, however, to sampling error.

The participating students had a mean age of 9.45 years ($SD = 0.88$) and included 11 second graders, 16 third graders, and 24 fourth graders. According to students' Individualized Education Programs (IEPs), students were diagnosed with learning disabilities ($n=21$; 41.18%), other health impairments ($n=11$; 21.57%), functional or developmental delay ($n=8$; 15.69%), speech/language impairments ($n=7$; 13.73%), and autism ($n=4$; 7.84%). Students had IEP goals related to reading (88.24% of students), math (56.86%), behavior or social/emotional learning (70.59%), and speech/language (17.65%). Participating teachers reported that 16 students (31.37%) also received services from school-based programs for English Language Learners. See Table 1 for participant demographics compared to the reported demographic data for participants in Jenkins and colleagues' study.

District-wide policies required teachers to collect PM assessment data for all students identified for special education. The district's research committee approved this study on the

condition that members of the research team would offer to conduct school-based progress monitoring (FastBridge; Christ et al., 2015) of participating students. All participating teachers accepted this offer. Therefore, teachers had access to school-based CBM data collected by research assistants (RAs), independent of data collected for the study itself. This context differs from the context of the original study, since Jenkins and colleagues explicitly reported that their eleven participating special education teachers were not using CBM in their teaching practices. Similar to Jenkins et al., I provided all teachers their participating students' study data, including each student's average weekly growth rate, at the conclusion of the study. As needed, I provided assistance in interpreting data.

Materials

Members of my research team administered a total of 42 PM passages (AIMSWeb; Shinn et al., n.d) to each participating student. We administered a random sequence of passages (determined by a random number generator) to each participant to minimize sequence effects. Because there were only 33 AIMSweb passages for each grade level (with the exception of first grade, which had only 23 passages), we readministered passages in the same randomized order assigned to each participant beginning in the 12th week of data collection (the third passage of the eighth week of data collection for students reading at a first grade instructional level).

Readministering the random order of passages differs from the procedure employed by Jenkins et al., who supplemented the 33 AIMSweb passages with nine PM passages from Edcheckup (www.edcheckup.com). I departed from the original procedures because, though the sets of passages from AIMSweb and Edcheckup were assigned the same grade-level text difficulty by the vendors, the passages across vendors were not necessarily functionally

equivalent (Jenkins et al., 2017). This pattern of questionable equivalency across passage sets has been confirmed by studies in which researchers objectively measured text difficulty and/or compared student reading rate on passages across CBM vendors (Ardoin & Christ, 2009; Ford, Missal, Hosp, & Kuhle, 2017). Dr. Jenkins reiterated this concern during a follow up phone conversation in the initial planning stages of this replication study (J. Jenkins, October 9, 2018, personal communication).

Readministering passages once students read through the entire set of AIMSweb passages at their respective instructional level increases the potential for practice effects (Jenkins, Zumeta, & Dupree, 2005). Practice effects may inflate oral reading rate on passages previously read compared to novel passages. However, the evidence suggests that practice effects are negligible given a 10-week interval between initial and follow up administration (Jenkins et al., 2005). Given this 10-week interval, only students reading at a first grade instructional level in this study would have the potential to show elevated practice effects.

Results from a *t*-test indicated that the mean difference between true growth slopes for students reading at a 2nd to 4th grade instructional level vs. 1st grade instructional level was not statistically significant ($t(49)=1.6027$, $p=0.1154$). I calculated the standardized mean difference effect size for these data, using an online effect size calculator (www.campbellcollaboration.org/escalc/html/EffectSizeCalculator-SMD2.php). There was an effect size of $d=0.515$ [-0.1227, 1.1526] between true growth for students reading at a 2nd to 4th vs. 1st grade instructional level. This suggests that, though the difference was not statistically significant, students reading at a 1st grade instructional level, on average, demonstrated poorer true growth than students reading at higher instructional levels. Given the potential elevated risk of practice effects for students

reading at a 1st grade level, these results suggest that, if anything, those students would have performed even more poorly compared to their peers reading at higher instructional levels.

Table 1. Student Demographics

	Current Study				Jenkins et al. (2017)			
	Mean	SD	*n*	%	Mean	SD	*n*	%
Age	9.45	0.88			NR	NR		
Grade	3.25	0.80			4.23	0.95		
2nd			11	21.57			1	1.79
3rd			16	31.37			11	19.64
4th			24	47.06			24	42.86
5th			-	-			14	25.00
6th			-	-			6	10.71
Instructional Reading Level	1.90	0.77			2.80	0.90		
1st			13	25.49			2	1.79
2nd			28	54.90			21	19.64
3rd			8	15.69			21	42.86
4th			2	3.92			10	25.00
5th			-	-			2	10.71
Gender								
Female			14	27.45			20	35.71
Ethnicity (*n*=46)								
Hispanic			20	43.48			NR	NR
Race (*n*=50)								
White			25	50.00			NR	NR
Black			23	46.00			NR	NR
Hispanic (*write in)*			6	12.00			NR	NR
Other			3	6.00			NR	NR
EL Services								
Receives EL Services			16	31.37			NR	NR
Disability								
LD			21	41.18			44	78.57
EBD			0	0.00			0	0.00
S/LI			7	13.73			0	0.00
OHI			11	21.57			6	10.71
F/DD			8	15.69			1	1.79
I/DD			4	7.84			5	8.93
IEP Goals								
Reading			45	88.24			NR	NR
Math			29	56.86			NR	NR
Behavior or SEL			36	70.59			NR	NR
Speech/Language			9	17.65			NR	NR
Median WRC								
Baseline	51.88	30.10			NR	NR		
Range: 3 to 133 WRC								
Final (Week 13)	61.51	33.39			NR	NR		
Range: 1 to 139 WRC								

Note: n=51 unless otherwise noted for current study. Original study had a final sample of 56 participants. NR = Not Reported; LD = Learning Disability; E/BD = Emotional/Behavioral Disorder; S/LI = Speech/Language Impairment; OHI = Other Health Impairment; F/DD = Functional/Developmental Delay; I/DD = Intellectual/Developmental Disability, EL = English Learner, IEP = Individualized Education Program, SEL = Socio/Emotional Learning, WRC = Words Read Correctly.

Procedures

Three hired RAs (two female) served as examiners for this study. The three RAs were graduate students in Child Studies, Human Development Counseling, and Economic Development. I trained all RAs in administering and scoring CBM. The 1.5-hr training included a written and verbal overview of the CBM protocols, supervised practice with these procedures, and an assessment check out, on which all RAs obtained 98% words read correctly (WRC) accuracy or greater.

Data collection began in the second week of January. This was later in the school year than the timeline of the original study, for which data collection occurred in the fall. During baseline week, RAs determined students' instructional reading level by identifying the set of passages on which students' median reading rate (i.e., WRC) fell between the 10th and 50th percentile for the grade level of the passage set, up to the students' actual grade level. RAs began administering passages at the grade level corresponding to participating teachers' estimates for each student's instructional reading level. Based on the student's performance, RAs increased or decreased the grade level of the probes as needed until students met the instructional level criterion. These data became the first three data points for each student. In all subsequent weeks, RAs administered passages at the students' instructional level, which is a recommended practice to ensure sensitivity to growth (NCII, n.d.). According to initial baseline data, thirteen students (25.49%) read at a first-grade level, 28 (54.90%) at second, eight (15.69%) at third, and two (3.92%) at a fourth-grade level. This process of identifying instructional level differs from the procedures used by Jenkins et al., who exclusively used teacher estimates of students' instructional reading level. I added this extra criterion to ensure that passages matched students'

needs based on objective data (rather than relying on teacher-report alone), thereby increasing the confidence that the CBMs would be appropriately sensitive to change.

RAs individually-administered the random probe order of assessments to participating students each week. RAs administered three CBM passages a week for 14 weeks (baseline week and 13 weeks of data collection thereafter) between January and April. This departs from Jenkins et al.'s testing procedures and the original data collection plan described in the OSF pre-registration. As planned, RAs would have administered three CBM passages at baseline, three a week for 11 weeks, and six passages for the final week. This procedural change was motivated by the limited windows of time we were allowed to assess students at each school, many of which overlapped across schools. It would not have been possible to administer six assessments to each student in the final week of data collection given the resource and time constraints.

I assigned each participating student a consistent testing day each week. RAs tested three students on Tuesdays, 23 students on Wednesdays, and 25 students on Thursdays. In the case that a student was absent on his or her assigned testing day, RAs returned on Friday for make-up assessments (Mondays during weeks without school on Friday). There was a one-week break from data collection during the district's spring break, which occurred between the ninth and tenth week of data collection (between Week 8 and 9 after baseline week). RAs administered the passages each week sequentially. Students read a student version of the passage and RAs recorded student responses on the examiner version of the passage, which included a word count along the margins of the text. RAs audio recorded each test administration.

Consistent with Jenkins et al.'s procedures, RAs told participants, "It's time for a short reading check. I'm using a timer to remind me how long I need to listen. When I say 'please begin' start reading here [*pointing to the first word of the passage*]. Your job is to do your best

reading. Do you have any questions? [*Pause*]. Okay, please begin" (Jenkins et al., 2017, p. 46). RAs began the timer when the student read the first word of the passage. Students read for 1 minute, as RAs recorded errors (i.e., mispronunciations, skipped words, and hesitations >3s). In the case of hesitations, RAs provided students the word after 3 seconds. RAs did not count self-corrections or insertions as errors. At the end of the minute, RAs noted the last word students read. Then, they administered the next CBM passage. Upon completing the administration of all passages for the week, RAs thanked the student and returned him or her to class. RAs recorded the total number of words the students read, the number of errors, and the WRC. They calculated the WRC by subtracting the number of errors from the total words read in the minute.

RAs scanned and uploaded all of their scored passages. A second scorer rescored each of these passages for inter-scorer reliability. I calculated inter-scorer agreement by calculating the percent of CBM passages with first and second scorer agreement on the WRC score, averaging across students and weeks. Inter-scorer reliability was high (96.08%; range by student: 80.95-100%). In instances of disagreement between the primary and secondary scorer, I served as third scorer and resolved the discrepancy through majority consensus.

RAs also conducted inter-observer reliability scoring on a planned, randomly selected set of 13 passages for each student. A second scorer independently listened to the audio from reliability assessments and scored student responses. Due to student absences and rare instances of audio files in which a student's voice was not captured adequately enough to score his or her reading, RAs actually completed inter-observer reliability scoring on 8 to 13 passages per student, accounting for 30.0% of all passages administered. This accounted for a greater proportion of reliability observations than initially planned for in the OSF pre-registration. Following the protocol described by Jenkins and colleagues, I calculated inter-observer

reliability by dividing the lower WRC score by the higher WRC scores from the lead and reliability data. I averaged these values across all reliability passages. Inter-observer agreement was high overall (*M*=97.18%), by passage (range: 75-100%), and by student (*M* range: 88.88-99.26%). The three instances in which agreement fell below 80% occurred with students who read fewer than 10 words correctly in a minute.

To control for the fact that the special education teachers had access to weekly school-based CBM data for participating students, I had participating teachers complete an initial survey of reading instruction for each participating student. On this survey, teachers indicated details related to students' reading instruction/intervention at baseline (e.g., session length and frequency, grouping type, and time dedicated to each area of reading instruction). I planned to have the teachers complete a weekly survey of instructional changes (see OSF pre-registration). However, after the start of data collection, I changed this procedure so that teachers completed a midpoint and final survey to determine the presence of any meaningful instructional changes (in instructional content, intervention dosage, or grouping) between survey time points. I made this adaptation from my pre-registered plan due to the inability of teachers to complete the surveys on a weekly basis. I used data from these surveys to determine the need to statistically control for instructional changes.

Design and Analysis

In the spirit intended by direct replication, I conducted identical primary data analyses as those employed by Jenkins and colleagues. Prior to conducting any statistical analyses, I cleaned the data by excluding data of any participant who missed more than one week of data collection. In Stata/ SE 14.0 (StataCorp, 2015), I used OLS regression to regress time on WRC scores to

calculate true growth slopes and all relevant weekly slopes. To account for each data point, I followed Jenkins and colleagues' procedure of using individual scores in all slope calculations, adding 0.003 days (5 min) to each additional measure administered in the same day. I compared all PM schedules and the true growth estimate to the goal growth rate used by Jenkins et al. (i.e., 1.0 WRC increase per week).

Growth estimates and PM schedules. I conducted an OLS regression using all 42 CBM data points to obtain a true growth estimate for each student. I also calculated weekly slopes (using OLS regression) for the same intermittent PM schedules analyzed by Jenkins and colleagues. Those six schedules included (a) one CBM weekly, using the first passage administered each week; (b) two CBMs every two weeks, using the first two passages administered in Weeks 2, 4, 6, 8, 10, and 12; (c) three CBMs every three weeks, using the three CBMs administered in Weeks 3, 6, 9, and 12; (d) three CBMs every four weeks, using the three CBMs administered in Weeks 4, 8, and 12; (e) three CBMs every five weeks, using the three CBMs administered in Weeks 5 and 10; and (f) three CBMs every six weeks, using the three CBMs administered in Weeks 6 and 12. This means that, for each student, I calculated a total of 12 weekly slopes for the weekly PM schedule, six weekly slopes for the Every-2 PM schedule, four weekly slopes for the Every-3 PM schedule, 3 weekly slopes for the Every-4 PM schedule, and two weekly slopes for both the Every-5 and Every-6 PM schedule. Additionally, I calculated weekly slopes for three other PM schedules. These included alternative versions of the weekly PM schedule (i.e., using the second or third CBM administered each week) and one CBM every two weeks (using the first passage administered in Weeks 2, 4, 6, 8, 10, and 12).

For PM schedules' weekly slopes, I ran the OLS regression with the available data that (a) been collected up to that point in time and (b) fit the respective PM schedule. In line with Jenkins and colleagues' procedures, I included the three baselines in calculating the weekly slopes for all intermittent PM schedules "to achieve a reliable estimate of baseline performance and ensure a common starting point" (Jenkins et al., 2017, p. 46). Table 2 illustrates the number of CBMs that contributed to each slope calculation for each PM schedule across the weeks of the study.

Table 2. Curriculum-Based Measurements per Week for Each Measured Slope

	True Growth	1 Every 1 Wk	2 Every 2 Wks	1 Every 2 Wks*	3 Every 3 Wks	3 Every 4 Wks	3 Every 5 Wks	3 Every 6 Wks
Baseline	3	3	3	3	3	3	3	3
Week 1	3	1 (4)						
Week 2	3	1 (5)	2 (5)	1 (4)				
Week 3	3	1 (6)			3 (6)			
Week 4	3	1 (7)	2 (7)	1 (5)		3 (6)		
Week 5	3	1 (8)					3 (6)	
Week 6	3	1 (9)	2 (9)	1 (6)	3 (9)			3 (6)
Week 7	3	1 (10)						
Week 8	3	1 (11)	2 (11)	1 (7)		3 (9)		
Week 9	3	1 (12)			3 (12)			
Week 10	3	1 (13)	2 (13)	1 (9)			3 (9)	
Week 11	3	1 (14)						
Week 12	3	1 (15)	2 (15)	1 (10)	3 (15)	3 (12)		3 (9)
Week 13	3 (42)							

Note. Parentheses show cumulative number of measures used to compute a slope at a given week. *Schedule not assessed by Jenkins et al. (2017).

Assessing adequacy of student growth. After conducting all OLS regressions, I assessed the adequacy of student growth. First, I assessed the adequacy of student growth as determined by true growth (OLS regression slope that took into account all 42 CBM probes). If a student's true growth slope met or exceeded the goal growth of 1.0 WRC per week, that student would have demonstrated adequate growth. If the student's true growth slope was less than 1.0 WRC increase per week, that student would have demonstrated inadequate growth. I created a dichotomized, "adequate growth" variable to indicate the adequacy of each student's true growth (1=adequate growth; 0=inadequate growth). Second, I assessed the adequacy of student growth, as determined by each weekly slope for all PM schedules. I created a dichotomized "adequate growth" variable for each weekly slope, indicating adequacy of student growth across weeks according to data from the PM schedules.

Decision accuracy. I compared the dichotomized "adequate growth" variable for each PM schedule's weekly slope against the dichotomized "adequate growth" variable for true growth, and determined whether those values matched or not. Matched decisions would have meant that either both the PM schedule's weekly slope and true growth determined adequate growth, or both determined inadequate growth. Unmatched decisions occurred when the PM schedule's weekly slope indicated adequate growth and true growth indicated inadequate growth, or vice versa. I created a dichotomized "decision match" variable based on this determination (1=decision match; 0=decision did not match). Finally, I determined decision accuracy by calculating the proportion of matched decisions for each PM schedule across participating students with data for the given week.

Additional data analyses. Similar to Jenkins et al., I ran additional data analyses in STATA (StataCorp, 2015) to supplement the primary analyses. First, I ran a binomial test (*di binomialtail[n,k,p]*, where *n*=number of students with data for the week, *k*=number of decision matches, and *p*= 50% chance of success) for each schedule's decision accuracy by week, to calculate whether obtaining each accuracy level or higher was significantly above chance (i.e., 50%). Second, I calculated the correlation between (a) true growth slopes and student grade level and (b) true growth slopes and student reading level to assess the relation between each of these variables and student true growth. Third, I calculated descriptive statistics to report the average true growth slopes, the standard deviation of those true growth slopes, and the skewness of the distribution of the true growth slopes across participants. Lastly, I calculated the number of participants failing to achieve the true growth goal rate of 1.0 WRC increase or greater per week across study weeks.

Supplementary data analyses. I conducted two supplementary analyses in STATA (StataCorp, 2015) that extended the work of Jenkins and colleagues. First, I ran point biserial correlations to determine whether there was a statistically significant correlation between teacher-reported instructional changes and students' true growth slope. This differs from the original pre-registered plan, given the change in data collection related to teacher-reported instructional changes (i.e., the switch to midpoint and final surveys, rather than weekly check ins), and the change to point biserial correlations (rather than a logit regression, which is best applied to data sets in which the outcome – not predictor – is categorical). I ran the point biserial correlations at the study's midpoint (i.e., after Week 6) and again at the study's conclusion (i.e., after Week 13). For the final week, I calculated two point-biserial correlations: (a) between

students' true growth slope and teacher-reported instructional changes between the midpoint and final survey, and (b) between students' true growth slope and teacher-reported instructional changes at any time in the study. In the event of a significant association between teacher-reported instructional changes and true growth slope, I planned to use teacher-reported instructional changes as a control variable in the primary OLS regression analyses.

In my OSF pre-registration, I planned to run a repeated measure (RM) ANOVA using PM schedule as the within-subjects factor and time to accuracy threshold (70% and 75% as two separate thresholds) as the dependent variable. If this test had been significant, I had planned to run follow-up post-hoc paired *t*-tests with adjusted *p* values. This test would have determined whether there was a statistically significant difference in the timeliness of each PM schedule obtaining those accuracy thresholds. However, due to the nature of the data set (i.e., time to accuracy threshold variable was a single value for each PM schedule), there was no way to compare group means and *SD*. Therefore, I did not run this additional analysis.

I also ran two additional exploratory post-hoc analyses. First, I ran a post-hoc correlation analysis with adjusted *p* values (using the Benjamini-Hochberg method of controlling for the false discovery rate; Benjamini & Hochberg, 1995) to determine whether the accuracy of the PM schedules (weekly, Every-2, Every-3, Every-4, Every-5, and Every-6) for each student was correlated with students' true growth slope. To conduct this analysis, I first averaged each PM schedule's "decision match" score by student across study weeks. Then, I calculated correlations between students' true growth and the average student-level decision accuracy for each PM schedule. The purpose of this analysis was to explore the relation between a PM schedule's accuracy for an individual student and that student's responsiveness. Second, I ran a post-hoc RM-ANOVA examining the main effects of PM schedule and time (week) on accuracy. In the

instance of a significant main effect of schedule, I planned to run follow-up post-hoc *t*-tests (with adjusted *p* values) comparing PM schedules' accuracy means. The purpose of this analysis was to explore how the main effect of PM schedule on accuracy statistically differed across PM schedules.

CHAPTER 3

RESULTS

In the following section, I describe the results from this study. As was the case with the methods section, I have written the results in a parallel format to Jenkins et al. to increase the ease of comparison across the original study and this replication study.

Over the 14-week period, the sample's mean true growth was 0.84 words per week (*SD*=0.55). This is less than Jenkins et al.'s reported sample true growth (*M*=1.12; *SD*=0.88). The distribution of true growth slopes across participants was approximately symmetric, with a nonsignificant skewness of 0.22. This is slightly more skewed than the distribution of Jenkins et al.'s data, which had a skewness of -0.05. While true growth slopes for 45% of Jenkins et al.'s sample was less than the goal rate of 1.0 WRC increase per week, 68.63% of the current sample failed to achieve the goal growth rate. Similar to Jenkins and colleagues' findings, true growth for this study's sample was not significantly correlated with grade (0.04, compared to Jenkins's correlation of -0.24) or reading level (0.17, compared to Jenkins's correlation of -0.23).

Table 3 summarizes the results from the teacher surveys about participating students' reading instruction (i.e., intervention context, grouping, and reading emphasis). Teachers reported providing students a reading intervention five days a week (session length *M*=45.34 min; *SD*=15.34 min) in small groups (*M*=4.14 students). Teachers reported that the majority of their instruction (*M*=70.80%) occurred using a small group format. Additionally, teachers reported that their instruction focused primarily on comprehension (*M*=29.75% of time), fluency (*M*=26.30% of time), and phonics-based instruction (*M*=22.61% of time).

Teachers reported changing the instruction (intervention context, grouping, or reading emphasis) for 16 students (31.37%) between the baseline and midpoint survey and nine students (17.65%) between midpoint and final survey. Of those students, teachers reported changing instruction for four students at both midpoint and final. In total, teachers reported changing the instruction of 21 students (41.18%) at some point in the study. One participating teacher went on medical leave between the midpoint and final survey. On the final survey, she reported the shift to a substitute teacher as the only instructional change for her participating students. This accounted for four of the students with reported instructional changes on the final survey, two of whom also had a reported instructional change on the midpoint survey. The results of the point biserial correlation tests indicated that there was not a significant relation between students' true growth and teacher-reported instructional changes between Weeks 1-6, Weeks 7-12, or across the entire study duration. Therefore, I did not control for instructional changes in the broader OLS regression analyses.

Table 3. Results from Teacher Survey about Students' Reading Intervention

Baseline Survey Items	Mean	SD
Intervention Context		
Session length (min)	45.34	15.34
Frequency (# of days/wk)	5.00	0.00
Group size	4.14	1.75
Time Spent across Grouping Types (%)		
Whole class	13.86	24.64
Small group	70.80	32.13
Partner work	4.20	6.23
Individual work	9.66	8.90
Time Spent on Each Big Idea in Reading (%)		
Phonological Awareness	8.86	10.78
Alphabetic Knowledge	1.75	3.24
Phonics	22.61	22.04
Fluency	26.30	22.82
Vocabulary	6.52	8.47
Comprehension	29.75	22.90
Writing	4.93	5.36

Teacher-Reported Changes to Students' Instruction	*n*	%
Change at Midpoint (Week 6)	16	31.37
Change Final (Week 12)	9	17.65
Change at Both Midpoint and Final	4	7.84
Change at Any Point in Study	21	41.18

Note. *n*=44. Three students received no reading intervention (special education consult-only). Missing data from four students.

Decision Accuracy

I calculated decision accuracy, following the procedure previously described, for PM schedules at each week of the study. Figures 2 through 4 show the decision accuracy of PM schedules across the weeks of the study. The accuracy of the weekly PM schedule (based on the first CBM given each week; Weekly [1st]) is represented in each of the figures as a comparison for each of the more intermittent PM schedules analyzed. Figure 2 compares all three simulated versions of the weekly PM schedule. Figure 3 compares the traditional weekly PM schedule (i.e., "Every week [1st probe]") with the two biweekly PM schedules (1-Every-2 and 2-Every-2). Finally, Figure 4 compares the weekly schedule with the Every-3, -4, -5, and -6 PM schedules.

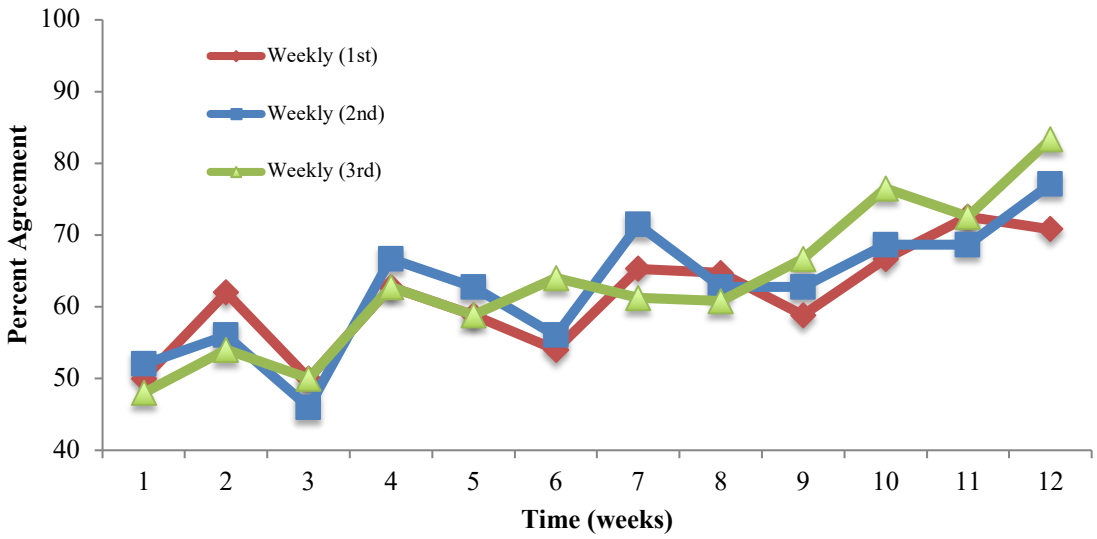

Figure 2. Decision accuracy of three weekly PM schedules.

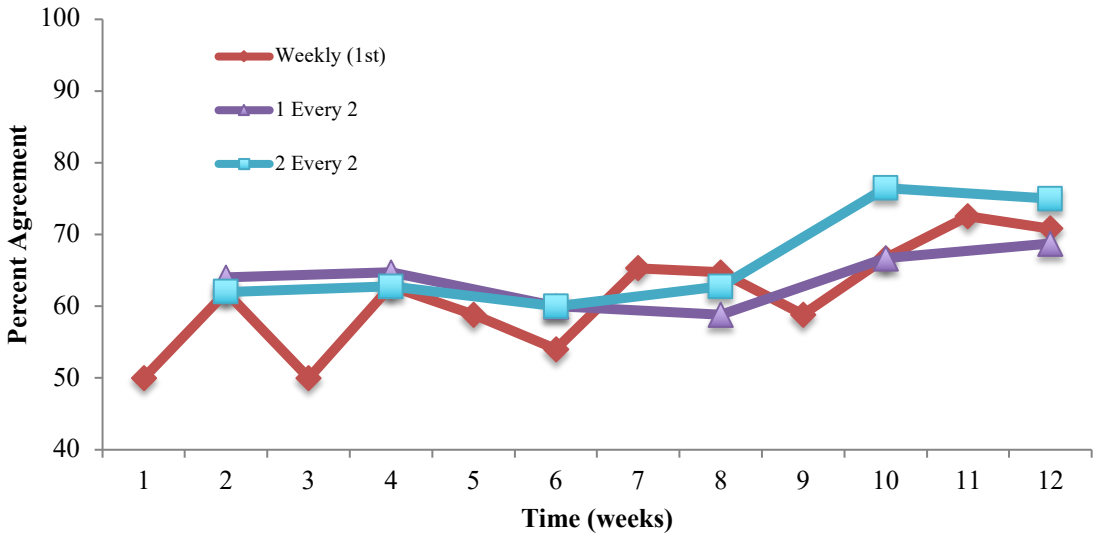

Figure 3. Decision accuracy of weekly vs. biweekly PM schedules.

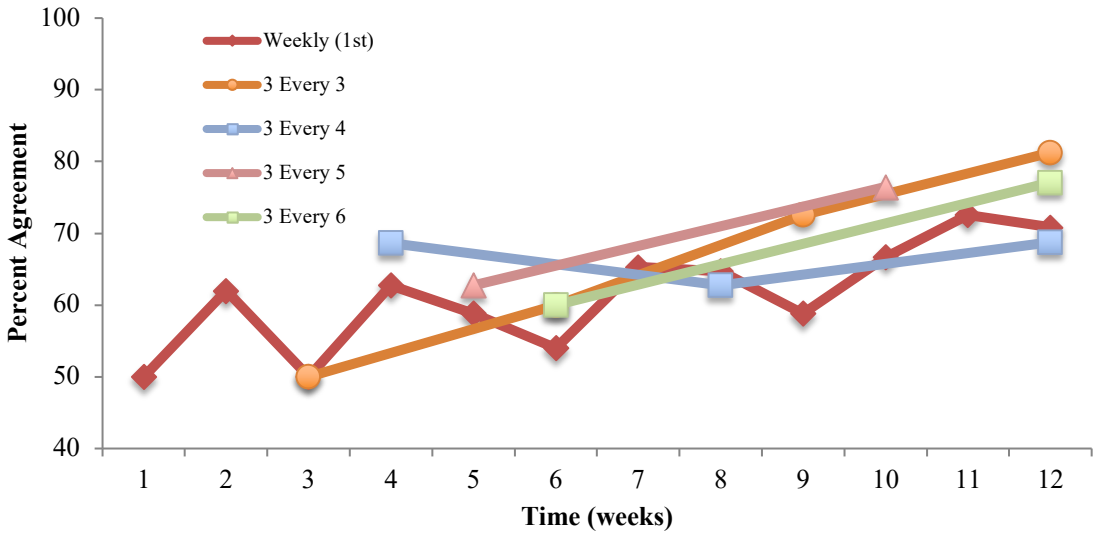

Figure 4. Decision accuracy of weekly vs. intermittent PM schedules.

For all PM schedules, decision accuracy increased across time, though imperfectly due to variability of accuracy across weeks for each PM schedule. For instance, the decision accuracy of the traditional weekly PM schedule decreased in weeks 3, 5, 6, 9, and 12 of data collection, compared to the week prior. Table 4 shows the decision accuracy of each schedule across the weeks of the study. Within each week, I sorted the PM schedules from most to least accurate and shaded gray the row for the traditional, weekly PM schedule. Table 4 also shows the overlap between the 39 true growth passages administered after baseline and the number of passages contributing to each PM schedule's weekly slope calculations. Finally, Table 4 shows the results of the binomial tests, which assessed whether obtaining each accuracy level or higher was significantly above chance (i.e., 50%). Similar to the results reported by Jenkins et al., Week 4 was the first week in which decision accuracy reached significance for three PM schedules, but the significance of the decision accuracy fluctuated across time. By Week 11, the decision accuracy of the weekly PM schedule reached a significance of $p>.01$. The following week, the decision accuracy of all PM schedules had reached a $p>.01$ significance level.

Table 4. PM Schemes for Decision Points: Most to Least Accurate

Decision Point: PM Schedule	**Accuracy (%)**	**Score Overlap (%)**	**Jenkins et al. (2017) Accuracy**
Deciding at Week 1			
Every week (2nd probe)	*52.0*	*2.6*	*N/A*
Every week (1st probe)	50.0	2.6	NR
Every week (3rd probe)	*48.0*	*2.6*	*N/A*
Deciding at Week 2			
1 every 2 weeks	*64.0**	*2.6*	*N/A*
Every week (1st probe)	62.0	5.1	NR
2 every 2 weeks	62.0	5.1	NR
Every week (2nd probe)	*56.0*	*5.1*	*N/A*
Every week (3rd probe)	*54.0*	*5.1*	*N/A*
Deciding at Week 3			
Every week (1st probe)	50.0	7.7	NR
Every week (3rd probe)	*50.0*	*7.7*	*N/A*
3 every 3 weeks	50.0	7.7	NR
Every week (2nd probe)	*46.0*	*7.7*	*N/A*
Deciding at Week 4			
3 every 4 weeks	68.6**	7.7	71.4**
Every week (2nd probe)	*66.7**	*10.3*	*N/A*
1 every 2 weeks	*64.7**	*5.1*	*N/A*
Every week (1st probe)	62.7*	10.3	64.3*
Every week (3rd probe)	*62.7**	*10.3*	*N/A*
2 every 2 weeks	62.7*	10.3	66.1*
Deciding at Week 5			
Every week (2nd probe)	*62.7**	*12.8*	*N/A*
3 every 5 weeks	62.7*	7.7	71.4**
Every week (1st probe)	58.8	12.8	58.9
Every week (3rd probe)	*58.8*	*12.8*	*N/A*
Deciding at Week 6			
Every week (3rd probe)	*64.0**	*15.4*	*N/A*
2 every 2 weeks	60.0	15.4	73.2**
1 every 2 weeks	*60.0*	*7.7*	*N/A*
3 every 3 weeks	60.0	15.4	76.8**
3 every 6 weeks	60.0	7.7	78.7**
Every week (2nd probe)	*56.0*	*15.4*	*N/A*
Every week (1st probe)	54.0	15.4	66.1*

Deciding at Week 7			
Every week (2nd probe)	*71.4***	*17.9*	*N/A*
Every week (1st probe)	65.3*	17.9	NR
Every week (3rd probe)	*61.2*	*17.9*	*N/A*
Deciding at Week 8			
Every week (1st probe)	64.7*	20.5	71.4**
Every week (2nd probe)	*62.7**	*20.5*	*N/A*
2 every 2 weeks	62.7*	20.5	73.2**
3 every 4 weeks	62.7*	15.4	67.9*
Every week (3rd probe)	*60.8*	*20.5*	*N/A*
1 every 2 weeks	*58.8*	*10.3*	*N/A*
Deciding at Week 9			
3 every 3 weeks	72.5**	23.1	76.8**
Every week (3rd probe)	*66.7**	*23.1*	*N/A*
Every week (2nd probe)	*62.7**	*23.1*	*N/A*
Every week (1st probe)	58.8	23.1	66.1*
Deciding at Week 10			
Every week (3rd probe)	*76.5***	*25.6*	*N/A*
2 every 2 weeks	76.5**	25.6	76.8**
3 every 5 weeks	76.5**	15.4	73.2**
Every week (2nd probe)	*68.6***	*25.6*	*N/A*
Every week (1st probe)	66.7*	25.6	75*
1 every 2 weeks	*66.7**	*12.8*	*N/A*
Deciding at Week 11			
Every week (1st probe)	72.5**	28.2	NR
Every week (3rd probe)	*72.5***	*28.2*	*N/A*
Every week (2nd probe)	*68.6***	*28.2*	*N/A*
Deciding at Week 12			
Every week (3rd probe)	*83.3***	*30.8*	*N/A*
3 every 3 weeks	81.3**	30.8	89.3**
Every week (2nd probe)	*77.1***	*30.8*	*N/A*
3 every 6 weeks	77.1**	15.4	83.9**
2 every 2 weeks	75.0**	30.8	83.9**
Every week (1st probe)	70.8**	30.8	78.6**
1 every 2 weeks	*68.8***	*15.4*	*N/A*
3 every 4 weeks	68.8**	23.1	83.9**

Note. Shaded area indicates the results of the traditional, weekly CBM schedule. PM = progress monitoring. Score overlap = Number of PM scores following baseline/true growth scores (*n*/39). Italicized PM schedules indicate additional schedules not evaluated by Jenkins et al. (2017).

*$p < .05$. ** $p < .01$. Binomial test; no correction for multiple tests.

Table 5 summarizes the ranking of PM schedules across study weeks. Contrasting the traditional weekly PM schedule (i.e., "Every week [1st probe]" in Table 2) with the intermittent PM schedules analyzed by Jenkins et al. (2-Every-2 weeks and 3-Every-3, -4, -5, and -6 weeks), a few patterns emerge. Nine weeks had both weekly and intermittent PM data. In Weeks 2 and 3, the weekly PM schedule had the same decision accuracy as the only intermittent schedule for the respective week. In Weeks 4, the weekly PM schedule had lower accuracy than the Every-3 schedule, but the same accuracy as the Every-2 schedule. In Weeks 5, 6, 9, and 10, the weekly PM schedule had the lowest accuracy of all relevant schedules for the week. In Week 8, the weekly PM schedule had the highest accuracy of all schedules for the week. Finally, in Week 12, the weekly PM schedule's accuracy was lower than the Every-2, Every-3, and Every-6 PM schedules, but higher than the Every-4 PM schedule. These results differ slightly from those reported by Jenkins et al., who, reporting patterns from Week 4 on, found that the weekly PM schedule was least accurate in five of the seven applicable weeks, and between the accuracy of two intermittent schedules in the other two weeks.

Table 5. Ranking of PM Schedules across Study Weeks

Week	Study	PM Schedule					
		Weekly	Every-2	Every-3	Every-4	Every-5	Every-6
Baseline		No contrasts					
1		No contrasts					
2	Current	*1st*	*1st*	-	-	-	-
	Jenkins et al.	NR					
3	Current	*1st*	-	*1st*	-	-	-
	Jenkins et al.	NR					
4	Current	*2nd*	*2nd*	-	1st	-	-
	Jenkins et al.	3rd	2nd	-	1st∗	-	-
5	Current	2nd	-	-	-	1st	-
	Jenkins et al.	2nd	-	-	-	1st∗	-
6	Current	4th	1st	*2nd*	-	-	*2nd*
	Jenkins et al.	4th	3rd∗	2nd∗	-	-	1st∗
7		No contrasts					
8	Current	1st	*2nd*	-	*2nd*	-	-
	Jenkins et al.	2nd∗	1st∗	-	3rd∗	-	-
9	Current	2nd	-	1st∗	-	-	-
	Jenkins et al.	2nd	-	1st∗	-	-	-
10	Current	3rd	*1st∗*	-	-	*1st∗*	-
	Jenkins et al.	2nd∗	1st∗	-	-	3rd∗	-
11		No contrasts					
12	Current	4th∗	3rd∗	1st∗	5th∗	-	2nd∗
	Jenkins et al.	5th∗	*2nd∗*	1st∗	*2nd∗*	-	*2nd∗*

Note. Weekly indicates the "Every week (1st probe)" PM schedule. Italicized text indicates tied accuracy value within a week. *Accuracy levels reached *a priori* accuracy threshold (70%). NR = Not Reported.

Jenkins et al. also reported that intermittent PM schedules were at least as accurate as the weekly PM schedule in 11 of the 15 contrasts from Week 4 on. They defined a contrast as each comparison between the weekly PM schedule and an intermittent PM assessed in the same week. Similar to the results reported by Jenkins et al., the results of this study indicated that intermittent PM schedules were at least as accurate in 12 of the 15 contrasts. Of those 15 contrasts, however, only seven included at least one of the comparison schedules reaching the minimum threshold of 70% accuracy. In six of the seven contrasts in which a schedule reached the 70% accuracy threshold, the intermittent schedules were more accurate than the weekly schedule. Only five contrasts included at least one of the comparison schedules reaching the 75% accuracy threshold, indicating that PM schedules for this population do not measure true growth well. In all five of these contrasts, the intermittent PM schedule was more accurate than the weekly schedule.

Descriptively, I examined the types of errors present in inaccurate decisions for PM schedules. Specifically, I tracked the instances in which a PM schedule misidentified a student whose true growth data showed inadequate growth (i.e., false positive, or Missed Non-Responder) compared to the instances in which a PM schedule misidentified a student whose true growth data showed adequate response (i.e., false negative, or Missed Responder). Figure 5 provides a count of the error types for each PM schedule across the study weeks. For nearly every week and PM schedule, the more common error type was the false positive error. This indicates a higher prevalence of Missed Non-Responders compared to Missed Responders.

The post-hoc, exploratory correlation analyses of the relation between student-level accuracy of the various PM schedules and each student's true growth showed that only the correlation between true growth and the 2-Every-2 PM schedule was significant (-0.3090; p=0.0274); however, this correlation did not remain significant after Benjamini-Hochberg

corrections, where the significance threshold dropped to $p>0.0083$. Correlations between true growth and other PM schedules (Weekly, Every-3, Every-4, Every-5, and Every-6) were all also negative in value – indicating that PM schedules had, on average, worse accuracy levels for students with slower rates of growth – but not significantly so. Furthermore, the post-hoc RM-ANOVA indicated a significant main effect of time (wk), but not of schedule, thereby eliminating the need for follow-up paired *t*-tests.

Finally, I ran the analyses on three additional PM schedules not explored by Jenkins and colleagues (i.e., Every week [2nd probe], Every week [3rd probe], and 1-Every-2-weeks). The three different simulated "weekly" PM schedules, which accounted for either the first, second, or third passage administered each week, demonstrated similar decision accuracy relative to each other. Each of the weekly PM schedules was most accurate relative to the other two weekly PM schedules in four of the 12 weeks. Additionally, the 1-every-2-weeks PM schedule, which was the schedule used by approximately half of the participating special education teachers for school-based PM assessments, was more accurate than the traditional, weekly PM schedule in three of the six applicable weeks, equally accurate in one week, and less accurate in two weeks.

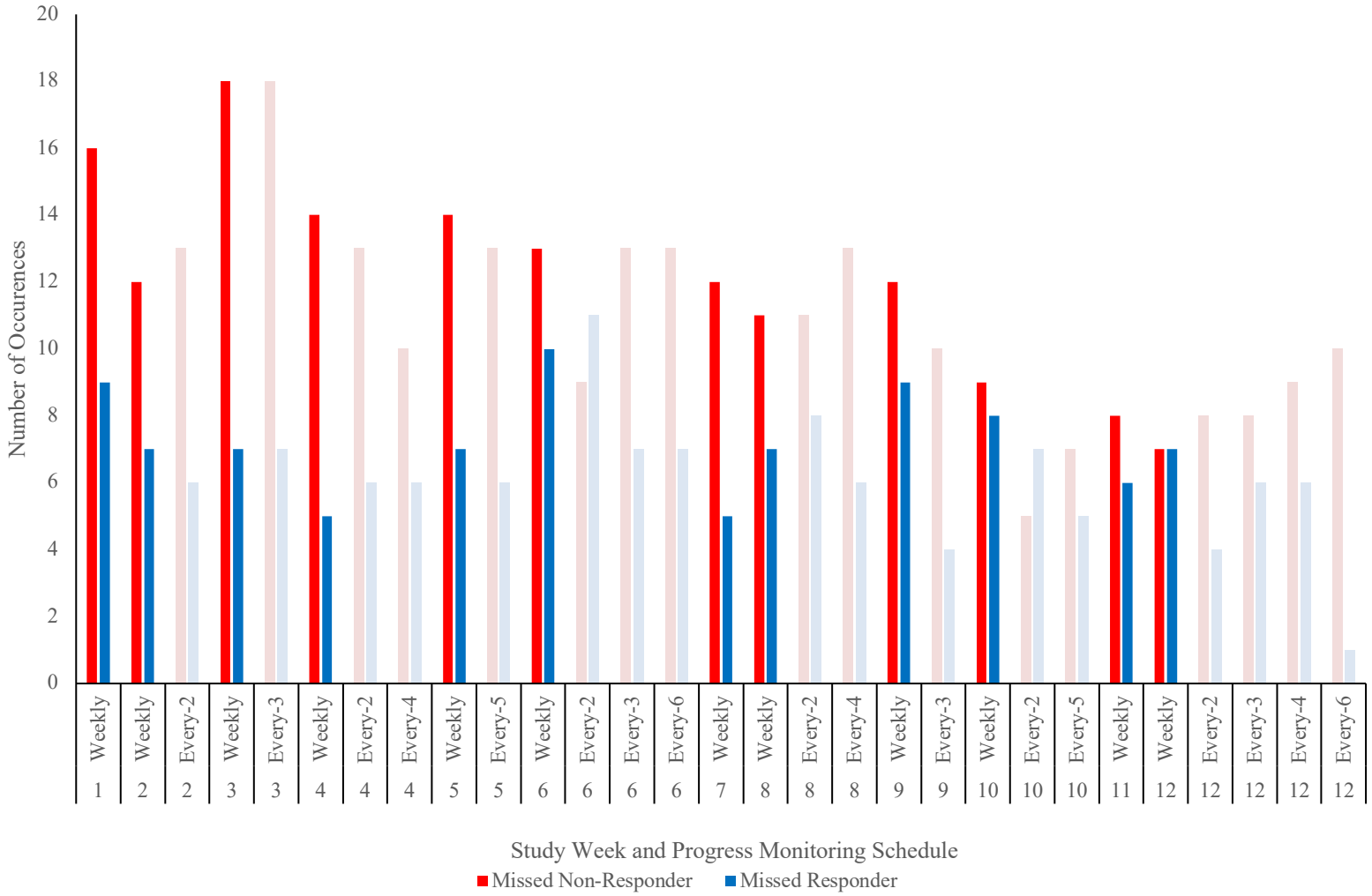

Figure 5. Count of error types (missed non-responder and missed responder) by schedule across study weeks. The bolder colors represent the data for the traditional, weekly PM schedule.

Timeliness

Table 6 shows the number of weeks it took each PM schedule to reach 70% and 75% accuracy the first time. Table 6 also shows how these results compare to those reported by Jenkins and colleagues. The Every-3 PM schedule reached 70% accuracy the first time the earliest (Week 9), and reached 75% the next time that schedule was assessed (Week 12). The Every-2 and Every-5 PM schedule reached 70% *and* 75% accuracy the first time in Week 10. The Every-2 PM schedule maintained 75% accuracy in Week 12. The traditional weekly (i.e., "Every week [1st probe]") and the Every-6 PM schedule reached 70% accuracy the first time in Week 11 and 12, respectively. The Every-6 PM schedule also reached 75% accuracy in Week 12. While the weekly PM schedule maintained accuracy levels above 70% in Week 12, it never reached the 75% accuracy threshold. The Every-4 PM schedule never reached either accuracy threshold. Across all PM schedules, the time to accuracy threshold with this sample was longer than the time reported by Jenkins et al. for their sample.

Table 6 also reports the time to accuracy thresholds for the additional PM schedules I analyzed in this study. The three different simulated "weekly" PM schedules required different amounts of time to reach accuracy thresholds the first time. While the weekly (1st probe) PM schedule reached 70% accuracy in Week 11 and never reached 75% accuracy, the weekly (2nd probe) reached both thresholds in Week 12 and the weekly (3rd probe) PM schedule reached both thresholds in Week 10. In Week 11, the weekly (3rd probe) PM schedule's accuracy decreased to 72.5%, but increased in Week 12 to 83.3%, the highest calculated accuracy. The 1-every-2-weeks PM schedule never reached either accuracy threshold.

Table 6. Time to Accuracy Thresholds

	Time to Accuracy Threshold (Wks)			
	70% Accuracy		75% Accuracy	
PM Schedule	Current	Jenkins et al.	Current	Jenkins et al.
Every week (1st probe)	11	8	Never	10
Every week (2nd probe)	12	N/A	12	N/A
Every week (3rd probe)	10	N/A	10	N/A
2 every 2 weeks	10	6	10	10
1 every 2 weeks	Never	N/A	Never	N/A
3 every 3 weeks	9	6	12	6
3 every 4 weeks	Never	4	Never	12
3 every 5 weeks	10	5	10	Never
3 every 6 weeks	12	6	12	6

CHAPTER 4

DISCUSSION

The purpose of this study was to replicate and extend the work of Jenkins and colleagues (2017). I explored two direct replication research questions: (a) "Is decision-making accuracy from intermittent PM inferior to that from weekly PM, the current standard?" (Jenkins et al., 2017, p. 45), and (b) "How many weeks of PM do these schedules require to reach specific levels of decision accuracy?" (Jenkins et al., 2017, p. 45). Additionally, I considered extension research questions to determine whether the comparison of PM schedules' decision accuracy differed when only considering schedules that had reached *a priori* accuracy thresholds, and whether intermittent PM schedules' timeliness was within two weeks of the timeliness of the weekly PM schedule. Finally, I aimed to consider whether the results of this study replicated the results reported by Jenkins and colleagues. I explored these questions with a sample of 51 students in 2nd through 4th grade students identified with high incidence disabilities. Overall, the results suggested that intermittent PM schedules had greater accuracy and better timeliness than weekly PM schedules in almost all incidences. Intermittent schedules met *a priori* accuracy thresholds, and therefore accurately reflected students' true growth, more often and quickly than weekly PM schedules. These results suggest non-inferiority of intermittent PM schedules compared to weekly PM, which replicates the conclusions asserted by Jenkins et al. (2017).

As with previous sections, I have paralleled the format of Jenkin et al.'s discussion section, for ease of comparison between the two studies. Within the sub-sections for each research question, I discuss the extent to which the current study's results replicated the results

reported by Jenkins and colleagues. I end by discussing limitations of this study and potential next steps in this line of research.

Does Intermittent PM Undermine Decision Accuracy?

In line with Jenkins and colleagues' findings, every PM schedule's decision accuracy increased with time. Mathematically, this is to be expected, considering the increased percent of score overlap of the data used to calculate PM schedules' weekly slopes and data used to calculate true growth across time (see Table 4). Intermittent PM schedules were at least as accurate as the traditional, weekly PM schedule in the majority of weeks and the majority of specific weekly vs. intermittent contrasts. These results are in line with my initial hypothesis that decision-making accuracy from intermittent PM would be indeterminately different from that of weekly PM. This provides preliminary evidence for the comparability of intermittent PM schedules compared to weekly PM.

There was a relatively small range in accuracy of PM schedules within a given week (e.g., of the schedules also evaluated by Jenkins et al., there was a 0 to 14.5 percentage point difference between the most to least accurate schedule in a week). Additionally, the post-hoc analyses indicated that PM schedules' student-level accuracy was not correlated with students' true growth, nor was there a significant main effect of schedule on mean accuracy of PM schedules. Further, like Jenkins and colleagues reported, the Every-3 PM schedule descriptively had either tied for or was the *most* accurate schedule across all relevant weeks (see Table 4 and Figure 3).

I extended the first research question by considering whether the comparability of weekly vs. intermittent PM schedules differed when only considering comparisons of schedules in which

at least one schedule met an *a priori* accuracy threshold of 70% and 75%. Fewer than half of the weekly vs. intermittent PM schedule contrasts (n=7) included at least one of the comparison schedules reaching the minimum 70% accuracy threshold. In all but one of those contrasts, the intermittent schedule was more accurate than the weekly schedule. Even fewer contrasts included at least one of the comparison schedules reaching the 75% accuracy threshold; however, in all five of those contrasts, the intermittent PM schedule was more accurate than the weekly schedule. While this is not a statistical test comparing PM schedules, it provides preliminary evidence that counters my hypothesis that decision-making accuracy from intermittent PM would be indeterminately different from that of weekly PM. Instead, these results suggest intermittent PM schedules may be more accurate than weekly PM schedules, when considering *a priori* accuracy thresholds.

These results may be driven by nature of the PM schedules themselves, since weekly PM schedules used only a single data point each week. Using only a single data point each week makes these data more sensitive to the fallibility of the assessment (e.g., variability in CBM passages and contextual differences between sessions) than PM schedules that aggregated multiple data points within a week (see Yoder et al., 2018, p. 56). It is possible that this effect of aggregating data points factors into the finding in both the current and original study that the Every-3 PM schedule – which accounted for the same number of passages as the weekly schedule (i.e., same score overlap with true growth) – was consistently more accurate than the weekly PM schedule.

Do students perform more poorly on initial passages administered? Jenkins and colleagues asserted that it is possible intermittent PM schedules outperformed the weekly PM

schedule because students may perform more poorly on initial passages administered in a week compared to later passages, which the traditional, weekly PM schedule would fail to capture. For this reason, I examined alternative "weekly" PM schedules using the second and third probe given each week. As can be seen in Figure 2, the three versions of the weekly PM schedule had comparable accuracy across all study weeks. More likely, the poorer accuracy of the weekly PM schedule (compared to intermittent PM schedules) relates to variability in CBM passages, the effect of which can be attenuated by aggregating data points around a given time point, such as what occurs with PM schedules that use a greater number of CBM passages at each time point, even when those time points are more spread out. This aligns with the principle of aggregation in classical measurement theory, which states that aggregating "a set of multiple measurements is a more stable estimator than any single measurement" (Yoder et al., 2018, p. 56).

How Many Weeks of PM Are Needed for Decision Making?

Overall, it took most PM schedules nine to 12 weeks to reach the 70 and 75% accuracy threshold explored by Jenkins and colleagues. This was 2-3 weeks longer than amount of time I hypothesized it would take PM schedules to reach each accuracy threshold. It was also a longer amount of time than Jenkins and colleagues reported it took PM schedules to reach the same accuracy thresholds for their sample. These results, however, should be couched in a broader discussion of whether these thresholds are the most appropriate or desirable thresholds to consider. For the purpose of direct replication, I used the same accuracy thresholds that Jenkins et al. used. Despite this, there is a need to explore the most "reasonable criterion" (Jenkins et al., 2017, p. 50) for sufficient accuracy required for data-based decision-making, such that special educators may be able to assess student response to interventions and make data-based decisions

as quickly as possible, while remaining confident that the data are reflecting students' true performance. Future research is needed to establish a greater evidence base for such accuracy threshold guidelines. "Timeliness" would vary depending on these guidelines.

I extended the second research question by considering whether the time it took intermittent PM schedules to reach each accuracy threshold was within two weeks of the time it took weekly PM schedules to reach the same accuracy threshold. The time it took intermittent PM schedules to reach the 70% accuracy threshold was less than or within two weeks of the time it took the weekly PM schedule to reach the same accuracy threshold in nearly every instance (see Table 6). The one exception was the Every-4 PM schedule. This schedule never reached 70% accuracy in the weeks of the study; however, the accuracy of the Every-4 PM schedule would not have been assessed again until Week 16, meaning that, even if it reached 70% accuracy by that point, it would have been more than two weeks beyond the time it took the weekly PM schedule to reach 70% accuracy, which occurred in Week 11.

The time it took intermittent PM schedules to reach the 75% accuracy threshold was also less than or within two weeks of the time it took the weekly PM schedule to reach the higher accuracy threshold in all instances where this was possible to assess (see Table 6). It was more challenging to compare schedules in this way for the higher accuracy threshold, however, because the weekly PM schedule never reached 75% accuracy. It took 10 weeks for the Every-2 and Every-5 PM schedules to reach 75% accuracy. Since the weekly had not reached 75% accuracy in Week 12, this finding supports the improved timeliness of these intermittent schedules compared to weekly PM. In fact, this result suggests superiority of these schedules' predictive properties compared to the weekly schedule. The Every-3 and Every-6 schedules reached 75% accuracy in Week 12, which was sooner than the weekly PM schedule. This

confirms the timeliness of these schedules compared to the weekly PM schedule; however, given that the weekly PM schedule never reached 75% accuracy by the end of data collection, I cannot assess the adequacy of the weekly PM schedule's timeliness compared to the intermittent schedules. Neither the Every-4 nor the weekly PM schedule had reached 75% accuracy by Week 12, making it impossible to assess the comparability of these schedules' timeliness in this way.

Do the Results of this Replication Study Replicate the Original Findings?

Jenkins and colleagues concluded that intermittent PM schedules were at least as accurate as weekly PM schedules across all weeks of the study. The results of this study replicated those initial findings. Additionally, Jenkins et al. found that it took intermittent PM schedules four to six weeks to reach 70% accuracy, and, for all intermittent PM schedules except the Every-5, six to 12 weeks to reach 75% accuracy. Jenkins et al. found that the weekly PM schedule took eight weeks to reach 70% and 10 weeks to reach 75% accuracy. Jenkins et al. report that this suggested little evidence of delayed decisions due to intermittent schedules.

In this replication study, PM schedules took longer than reported by Jenkins et al. to reach accuracy thresholds (by more than two weeks) in nearly all instances (see Table 6). Despite this, the results of this study similarly suggest little evidence of delayed decisions due to intermittent schedules, if timeliness is defined as the number of weeks it takes PM schedules to reach accuracy thresholds. These interpretations do not change when considering teacher-reported instructional changes, since the results of the point-biserial correlation tests indicated that I did not need to account for these changes in my analyses. It is important, however, to consider alternative definitions of timeliness (e.g., the time it takes a PM schedule to appropriately identify a student in need of instructional adaptations), which may provide a more

nuanced view of PM schedules and the role they play in the DBI process. I discuss this alternative definition of timeliness more completely in the "Next Steps" subsection of this discussion (p. 55).

How Does This Replication Study Compare to Jenkins et al.'s Study?

There were a few differences between the current study and the original study conducted by Jenkins and colleagues. First, there were dissimilarities in the sample that are important to note. Despite best recruitment efforts, my final sample was slightly smaller than Jenkins et al.'s final sample (51 students vs. 56 students), despite recruiting 64 students initially (compared to Jenkins and colleagues' initial sample of 66 students). The sample of students recruited for this study consisted of students from transient families with histories of frequent moves, students who demonstrated chronic absenteeism (e.g., missing more than 1 week of data collection despite make-up assessment procedures), or who experienced instability in home life (e.g., being put into foster care). As a result, there was a higher attrition rate in this study than in Jenkins and colleagues' study (20.31% of the initial sample was not included in the final sample, compared to 15.15% attrition reported by Jenkins et al.). These factors also potentially relate to the greater proportion of students who missed 1 week of data collection in this sample compared to the original study's final sample (17.65% vs. 8.93%). While the *t*-test result indicated that the difference in true growth for students with incomplete vs. complete data was not significant, there was a moderate effect size (d=-0.5069). With a larger sample size, this analysis would have greater power to detect group differences, and may indicate a significant difference between these groups of students.

Additionally, I targeted recruitment in local elementary schools. Given the grade level structure of the schools in the district, the current sample had a lower average grade level (3.25 vs. 4.23) and instructional reading level (1.90 vs. 2.80) compared to the sample data reported by Jenkins and colleagues. Further, the students in my sample were identified with a more diverse range of disabilities that the disabilities of the student participants in Jenkins and colleague's study (see Table 1). Jenkins and colleagues did not report demographic details such as students' race/ethnicity, or participation in additional English Language Learner-related services. Because of this, it is not possible to compare the samples across these domains.

There were also differences in the results of both studies. First, in all but one instance (i.e., Every-5 schedule during Week 10), Jenkins and colleagues reported higher decision accuracy for PM schedules than the calculated accuracy of the PM schedules for the current sample's data (see Table 4). This contributed to the greater statistical significance of the binomial tests Jenkins et al. conducted compared to the results of the binomial tests for the current study's data. It also contributed to the increased time it took each schedule to reach the 70 and 75% accuracy thresholds (see Table 6) for the current study. Second, a larger proportion of the current sample (68.63% compared to 45% of Jenkins and colleagues' sample) failed to achieve the goal rate of growth. This greater proportion of inadequate response is also reflected in the mean true growth rate for this sample (M=0.84; SD=0.55) compared to the mean true growth rate for the sample reported by Jenkins et al. (M=1.12; SD=0.88).

Jenkins et al. (2017) argue that their "results hint at the amount of PM needed to satisfy various accuracy criteria and provide a beginning database for guideline development" (p. 50) related to sufficient accuracy thresholds for PM schedules. The differences between the original study's results and the results of this current study bring to bear additional questions regarding

the extent to which guidelines may need to be calibrated differently for different samples of students. It is possible that underlying base rates, or prevalence, of inadequate response may contribute to the overall accuracy of PM schedules. This could be explored further through CBM demonstration studies where base rate could be manipulated across larger samples than is possible in typical special education research. Such research would also provide the opportunity to explore the sensitivity, specificity, positive predictive value, and negative predictive value of different PM schedules, thereby deepening the understanding of each PM schedule's diagnostic abilities.

Despite the larger proportion of inadequately responding students in the current sample, the participating special educators reported relatively few instructional changes for students at the midpoint (n=16; 31.37%) and conclusion (n=9; 17.65%) of the study. These preliminary results are in line with previous evidence that suggests teachers do not adequately use CBM data to inform instruction even in the best of circumstances, such as when RAs conduct the CBMs and/or CBM software provides instructional recommendations for adaptations (Stecker et al., 2005). Furthermore, there was not a significant correlation between the true growth of students in this sample and teacher-reported instructional changes. Considering the fact that my RAs also conduct school-based PM for each participating student – meaning teachers' time did not have to be dedicated to CBM administration and they could simply access student PM data collected for them regularly – the results of these instructional surveys bring to light questions related to the true nature of data collection time as a barrier to data-based decision-making in practice.

Limitations

While the results of this current study and the original study indicate intermittent PM schedules are indeterminately different from the traditional, weekly PM schedule, there are limitations to this study that could potentially impact the generalizability of the results. Given the decision to closely replicate the data collection procedures reported in the original study, many of these limitations align with those described by Jenkins and colleagues. Other limitations arose from aspects specific to this study's methodology and results. I describe each of these limitations in this section.

First, the duration of the study required 42 CBM passages, which exceeded the total number of available AIMSweb passages. After consulting with Dr. Jenkins, I chose to repeat the randomized order of passages once students read through the full set of available passages at their instructional level. While this raises the question of the potential for practice effects, previous research suggests this effect is diminished after 10 weeks (Jenkins et al., 2005). Nearly three-quarters of the sample read at a second through fourth grade instructional level, for which there were 33 available AIMSWeb passages. This meant that these students did not begin repeating passages until 12 weeks had passed. For the 25.49% of participating students reading at a first-grade instructional level, however, repeated reading of passages began in the third passage of the eighth week, since there were only 23 AIMSWeb passages for this grade level. The *t*-test results indicated that true growth for these students was not significantly different than the true growth for students reading at a higher instructional level. Though not significant, the effect size calculation showed that, on average, students reading at a 1st grade level actually demonstrated poorer true growth than students reading at a 2nd to 4th grade instructional level. Together, these results provide preliminary evidence that practice effect may not be playing a meaningful role in

scores of students repeating first grade passages in the eighth week of data collection. However, it is still important to note this as a potential threat to internal validity.

Second, as Jenkins and colleagues described, there is a limitation in the use of the same assessment data to estimate true growth and weekly slopes for each PM schedule. As the score overlap increases (i.e., as a PM schedule shares a greater proportion of CBM passages from the collective, "true growth" set), there would automatically be greater accuracy. This makes it challenging to ascertain what proportion of the variance of each PM schedule's accuracy should be attributed to score overlap and what proportion should be attributed to the diagnostic adequacy of the schedule itself. Using a completely independent set of passages to estimate true growth may be preferable, since this would eliminate the issue of score overlap. I was not able to use a separate set of passages, since this would have doubled the number of CBM passages administered to each participating student each week, and I was limited by scheduling constraints. Further, using a different set of passages to estimate true growth introduces the additional question of equivalency of CBM passages across vendors (see Ardoin & Christ, 2009; Ford et al., 2017) and the extent to which student growth on passages from one vendor is comparable to growth on passages from another vendor.

Third, there were recruitment, attrition, and student attendance issues that impacted the final sample size in this study. The smaller sample size, particularly relative to Jenkins and colleagues sample size, impacts the generalizability of the findings for this first cohort of participants. All results, therefore, should be considered preliminary. Additionally, Jenkins et al. (2017) did not report the characteristics of the sample they recruited. This makes it challenging to draw conclusions about the findings of this study compared to those of Jenkins and colleague's study.

Finally, some of the PM schedules never reached the 70% accuracy threshold in the 12 weeks of data collection. In part, this is due to the relatively lower decision accuracy of the PM schedules for the current sample of students, compared to the decision accuracy of PM schedules for students in the original study. Extending the number of weeks of data collection would allow for consideration of PM schedule timeliness more completely.

Next Steps

Because of the limitations of this study, I caution against making broad assertions that special educators should adopt intermittent PM schedules to ease the burden of assessment time. Future research is needed to explore these research questions further. In this section, I describe five potential paths for this future research that will help advance the field's understanding of the role PM schedule plays in data-based decision-making.

First, I plan to recruit a second cohort of students next year. This will address the sample size limitation and provide greater confidence in the generalizability of the results of this study. With this second cohort, I will also be able to begin to examine the potential relation between underlying prevalence of inadequate response and PM schedules' decision accuracy. The results of this study indicate lower accuracy across PM schedules relative to the accuracy of PM schedules reported by Jenkins and colleagues. At the same time, the sample for this study demonstrated higher rates of inadequate growth. I believe exploring the effect of prevalence is an important line of research that will allow for a more nuanced understanding of PM schedules' adequacy in identifying student growth. This is especially important considering that the large majority of errors of PM schedules in this study were false positives, meaning that PM schedules were more likely to miss non-responders (see Figure 5). This error has potentially important

implications for practice, since this would mean teachers engaging in data-based instructional decisions regularly would not have the necessary information to determine a need for an instructional adaptation. This would mean that the teachers of "Missed Non-Responders" would continue with instruction that is not adequately individualized to the inadequately responding student's instructional need, thereby perpetuating the inadequate response longer than necessary. Of the two types of potential errors, the risk associated with missing non-responders is higher.

Second, current DBI decision rules for instructional adaptations have moved beyond a purely growth rate-based metric. More often, current decision rules for making instructional adaptations include a metric that assesses student growth by considering the number of consecutive points a student's CBM performance falls below the expected performance level of the student's goal line. Future research could consider this alternative, decision rule metric for identifying student response, and use this metric to calculate decision accuracy across PM schedules. This would help determine whether the conclusions about decision accuracy across PM schedules replicate across alternative metrics for determining student response to intervention. I plan to re-analyze the current study data using this points-below metric to begin preliminary work in this area.

Third, this investigation examined the accuracy of oral reading fluency CBMs. Future research should include whether the results replicate across other reading CBMs (e.g., phoneme segmentation fluency, nonsense word fluency, word reading fluency) or other academic domains (e.g., math). Future studies in this area would provide a comprehensive view of PM schedule accuracy, independent of the specific skill assessed, and would help determine whether recommendations for PM schedule adoption in schools should differ depending on the target skills assessed. Additionally, future research should consider the comparability of students' true

growth across areas of reading and provide guidance to teachers whose students may demonstrate inconsistent patterns of growth across CBM types. For example, how should a teacher proceed if a students' true growth in oral reading fluency indicates inadequate response, but nonsense word reading fluency indicates adequate response? How should a teacher's plan change for a different student profile (e.g., a student demonstrating adequate oral reading fluency growth, but inadequate nonsense word reading fluency growth)? Future research should seek to help teachers prioritize making informed decisions about which CBMs to use and how to interpret potentially incongruent interpretations about the adequacy of student growth.

Fourth, future research should consider alternative definitions of "timeliness". While the work in this study and Jenkins and colleagues' study defined "timeliness" as the number of weeks it took different PM schedules to reach decision accuracy thresholds, another definition of "timeliness" would be one that more closely aligns with the DBI goal of identifying inadequate response to inform instructional changes (Danielson & Rosenquist, 2014). Namely, it is possible that a more meaningful metric of "timeliness" is the amount of time it takes PM schedules to identify inadequately responding students, based on true growth's determination of inadequate response. This alternative metric of "timeliness" would address the issue related to "Missed Non-Responders". Research in this area would support the field's understanding of the false positive errors in PM schedules decision accuracy. Furthermore, defining "timeliness" in this way could serve as an index for the decision-making discrepancy between different PM schedules. Since the decision-making discrepancy could be due to the fact that the various PM schedules collect data during different weeks of the study, it is possible that decisions to make instructional changes could be delayed in intermittent PM compared to weekly PM. Therefore, this alternative

"timeliness" index would be one way to capture an important aspect of CBM data collection – the use of available data to make *timely* instructional changes, especially for non-responders.

Finally, at the conclusion of this study, I interviewed all participating special education teachers. I asked questions about their practices related to data collection, evaluation, and application. In the future, I plan to transcribe these interviews and code for underlying themes across teachers. I will use data from these interviews to inform future research studies aimed at coaching teachers in the frequency and efficiency with which they engage in these practices, with and without researcher support. There is strong evidence supporting the use of CBM and DBI practices to improve student outcomes (Jung et al., 2018; Stecker et al., 2005), yet teachers demonstrate poor use of these practices overall (Stecker et al., 2005). Given this context, there is a need for future research to explore ways to improve teacher knowledge, skill, and continued use of data-based decision-making practices, in an effort to decrease the research to practice gap so that student outcomes may begin to improve.

Conclusion

The goal of this study was to replicate and extend the work of Jenkins et al. (2017). Given current initiatives related to expanding upon data-based decision-making frameworks in schools (see Lemons et al., 2019), the work of this replication study is important and has the potential to make an impact in the field of special education. While the current results are only preliminary, there is beginning evidence of replicated findings related to decision accuracy of different PM schedules. Though the aim of the original study was to consider intermittent PM schedules as a way to increase the feasibility of CBM administration and address the commonly reported barrier of time, the current results preliminarily suggest a more complicated reality

related to data use in practice, particularly for inadequately responding students. Regardless, the work in this study provides an important empirical rationale for future investigations into PM schedules. Such work could serve as a foundation for the future development of teacher-level interventions aimed at improving the inadequate prevalence of data-based decision-making in schools today. It is only through addressing these issues that we, as a field, may be able to alter teacher behaviors and, consequently, improve academic and life outcomes for students with the most persistent reading difficulties.

REFERENCES

Ardoin, S. P., & Christ, T. J. (2009). Curriculum-based measurement of oral reading: Standard errors associated with progress monitoring outcomes from DIBELS, AIMSweb, and an experimental passage set. *School Psychology Review, 38*(2), 266-283. Retrieved from http://login.proxy.library.vanderbilt.edu/login?url=https://search-proquest-com.proxy.library.vanderbilt.edu/docview/219656384?accountid=14816

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B, 57*, 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Christ, T. J. (2006). Short term estimates of growth using curriculum-based measurement of oral reading fluency: Estimates of standard error of the slope to construct confidence intervals. *School Psychology Review, 35,* 128–133.

Christ, T. J., Arañas, Y. A., Johnson, L., Kember, J. M., Kilgus, S., Kiss, A. J., ... Windram, H. (2015). Formative Assessment System for Teachers: Abbreviated Technical Manual for Iowa Version 2.0, Minneapolis, MN: Author and FastBridge Learning (www.fastbridge.org).

Christ, T. J., Zopluoglu, C., Monaghen, B. D., & Van Norman, E. R. (2013). Curriculum-based measurement of oral reading: Multi-study evaluation of schedule, duration, and dataset quality on progress monitoring outcomes. *Journal of School Psychology, 51*, 19–57. http://dx.doi.org/10.1016/j.jsp.2012.11.001.

Coyne, M., Cook, B. G., & Therrien, W. J. (2016). Recommendations for replication research in special education: A framework of systematic, conceptual replications. *Remedial and Special Education, 37*, 244-253.

Danielson, L., & Rosenquist, C. (2014). Introduction to the tec special issue on data-based individualization. *Teaching Exceptional Children, 46*(4), 6-12.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*(3), 219-232.

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education, 37*(3), 184-192.

Deno, S. L. (2014). Reflections on progress monitoring and data-based intervention. *Special education past, present, and future: Perspectives from the field* (pp. 171-194): Emerald Group Publishing Limited.

Deno, S. L., Fuchs, L., Marston, D., & Shin, J. (2001). Using curriculum-based measurements to establish growth standards for students with learning disabilities. *School Psychology Review*, *30*, 507–524.

Deno, S. L., & Mirkin, P. K. (1977). *Data-based program modification: A manual*. Reston VA: Council for Exceptional Children.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies in Reading, 5,* 239–256. http://dx.doi.org/10.1207/s1532799xssr0503_3.

Ford, J. W., Missall, K. N., Hosp, J. L., & Kuhle, J. L. (2017). Examining oral passage reading rate across three curriculum-based measurement tools for predicting grade-level

proficiency.*School Psychology Review, 46*(4), 363-378. Retrieved from http://login.proxy.library.vanderbilt.edu/login?url=https://search-proquest-com.proxy.library.vanderbilt.edu/docview/2087743678?accountid=14816

Gersten, R., Compton, D., Connor, C. M., Dimino, J., Santoro, L., Linan-Thompson, S., & Tilly, D. W. (2008). *Assisting students struggling with reading: Response to intervention and multi-tier intervention for reading in the primary grades. A practice guide* (NCEE 2009-4045). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc/ publications/practiceguides/

Individuals with Disabilities Education Improvement Act, Pub. L. No. 108-446 (2004).

Jenkins, J., Schulze, M., Marti, A., & Harbaugh, A. G. (2017). Curriculum-based measurement of reading growth: Weekly versus intermittent progress monitoring. *Exceptional Children, 84*(1), 42-54. doi:http://dx.doi.org/10.1177/0014402917708216

Jenkins, J., Zumeta, R., & Dupree, O. (2005). Measuring gains in reading ability with passage reading fluency. *Learning Disabilities Research & Practice*, *20*(4), 245–253. https://doi.org/10.1111/j.1540-5826.2005.00140.x

Jung, P. G., McMaster, K. L., Kunkel, A., Shin, J., & Stecker, P. M. (2018). Effects of data-based individualization for students with intensive learning needs: A meta-analysis. *Learning Disabilities Research & Practice, 33*, 144-155. doi:10.1111/ldrp.12172

Lemons, C. J., King, S. A., Davidson, K. A., Berryessa, T. L., & Gajjar, S. A. (2016). An inadvertent concurrent replication: Same roadmap, different journey. *Remedial and Special Education, 37*, 213-222.

Lemons, C. J., Sinclair, A. C., Gesel, S. A., Gandhi, A. G., & Danielson, L. (2019). Integrating intensive intervention into special education services: Guidance for special education administrators. *Journal of Special Education Leadership, 32*, 29-38.

National Center on Intensive Intervention. (n.d.). *Using academic progress monitoring for individualized instructional planning*. Retrieved from http://www.intensiveintervention.org/sites/default/files/Academic_Progress_ Monitoring-updated.pdf

Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., & Long, J. D. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlation evidence. *Journal of School Psychology, 47*, 427–469. http://dx.doi.org/10.1016/j.jsp.2009.07.001.

Shinn, M. R., Shinn, M. M., & Langell, L. A. (n.d.). Overview of curriculum-based measurement (cbm) and aimsweb. *Retrieved from http:// www.AIMSweb.com*.

StataCorp. (2015). *Stata statistical software: Release 14*. College Station, TX: StataCorp LP.

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools, 42*(8), 795-819. doi:http://dx.doi.org/10.1002/pits.20113

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, & National Assessment of Educational Progress. (2017). *2017 reading assessment*. Washington, DC: Author.

Yoder, P. J., Lloyd, B. P., & Symons, F. J. (2018). *Observational measurement of behavior (2nd ed.).* Baltimore, MD: Paul H. Brookes Publishing Co.