


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

Exceptional Children  
2020, Vol. 87(1) 92–112  
© The Author(s) 2020  
DOI: 10.1177/0014402920924845  
journals.sagepub.com/home/ecx  




Samantha A. Gesel<sup>1</sup>  and Christopher J. Lemons<sup>2</sup>

## Abstract

Curriculum-based measurement (CBM) is a systematic, ongoing assessment framework that allows special educators to monitor students' progress and determine the need for instructional adaptations. Jenkins and colleagues examined the accuracy and timeliness of six different schedules of CBM progress monitoring (PM). The authors found that weekly and intermittent PM schedules were similarly accurate and timely. This study replicated and extended the work of Jenkins and colleagues by examining the accuracy and timeliness of different PM schedules for 51 students with disabilities. Results indicated that the accuracy and timeliness of the PM schedules for the current sample was poorer than the accuracy and timeliness reported by Jenkins and colleagues. In line with the results of the original study, however, these results indicated that intermittent PM schedules sufficiently predicted student true growth compared to weekly PM schedule. Implications for research and practice are discussed.

Ongoing monitoring of students' progress toward annual goals and objectives is a core feature of special education (Individuals With Disabilities Education Act [IDEA], 2006). Curriculum-based measurement (CBM; Deno, 1985) is a common progress-monitoring (PM) framework used for this purpose. CBM includes brief, general outcome measures that assess student performance on academic skills through the use of multiple, equated probes (Deno, 2003). To monitor reading progress, teachers of students who are reading at a first- to eighth-grade instructional level often use an oral reading fluency (ORF) CBM, given the strong relation between the number of words read correctly (WRC) and reading achievement (Reschly et al., 2009). CBM can be used within the process of data-based individualization (DBI), a dynamic approach to intensive intervention in which teachers evaluate data to inform instruction (National Center on Intensive Intervention [NCII], n.d.). With DBI, teachers use data to

determine the adequacy of student progress toward annual goals and objectives and systematically adapt instruction when insufficient progress is demonstrated.

The process of monitoring students' progress toward annual goals and objectives is necessary and legally mandated (IDEA, 2006). Although the What Works Clearinghouse Practice Guide for Response to Intervention reported minimal evidence to support progress monitoring (Gersten et al., 2008), other reviews have provided evidence that the use of CBM and DBI practices positively impacts the academic growth of students with disabilities. In a narrative review, Stecker

---

<sup>1</sup>University of North Carolina at Charlotte

<sup>2</sup>Stanford University

## Corresponding Author:

Samantha A. Gesel, Department of Special Education and Child Development, University of North Carolina at Charlotte, 9201 University City Blvd., Charlotte, NC 28223.

Email: sgesel1@uncc.edu

et al. (2005) reported that CBM-based interventions resulted in positive effects on students' achievement in reading, mathematics, and spelling. In a more recent meta-analysis, Jung et al. (2018) reported significant positive student outcomes for interventions in which teachers individualized instruction based on CBM data alone (classified by the authors as "DBI Only";  $g = 0.37$ ) and CBM data combined with additional information or recommendations (classified as "DBI Plus";  $g = 0.38$ ). These reviews provide support for using CBM data in the iterative DBI process to improve learning outcomes of students with the most persistent needs.

NCII (n.d.) recommends that teachers collect PM data weekly when implementing DBI. However, special education teachers may consider alternative PM schedules to monitor their students' progress. Deciding how frequently to collect CBM data is a complex issue. PM schedules can differ in how frequently probes are administered and in the number of CBM probes administered at each data collection time point. Both of these factors, combined with the length of time across which data are collected, affect the accuracy with which CBM data represent a student's true performance (Christ et al., 2013). Christ and colleagues (2013) suggest that a minimum of 8 to 12 weeks of data is necessary to accurately use CBM for instructional decision making, depending on the quality of the data set (i.e., the size of the error residuals of the data).

Teachers must select PM schedules that are sufficiently accurate while considering the timeliness of instructional adaptations for students who are not responding adequately (Jenkins et al., 2017). Teachers may be able to make instructional changes for students more rapidly if they use fewer than 8 to 12 weeks of CBM data; however, they will be doing so with less reliable information, thus increasing the probability of inaccurate decisions. In contrast, there is greater confidence that the data-based decisions correspond to data reflecting students' true growth (TG) when teachers collect additional weeks of CBM data, but those additional weeks may extend the time students continue with instruction

that may not be adequately intensified or individualized for them. Although this trade-off between accuracy and timeliness is inherent in all CBM data collection, additional research is needed to examine how these factors vary across PM schedules.

Research in this area is especially relevant, given that teachers struggle to collect CBM data at all, let alone to engage in data-based decision making related to their instruction or instructional goals (Deno, 2003; Stecker et al., 2005). Often, teachers cite the need for dedicated time in their schedules to administer CBM assessments and evaluate student data to create a system in which DBI can be successful (Lemons et al., 2019). Without such a system, teachers often view time as the primary barrier to their use of CBM to inform instructional decisions (Deno, 2003). Given this reported barrier, there is a need to research ways to balance (a) teachers' time to administer CBM probes and use those data to adapt instruction with (b) selecting PM schedules that enhance the technical adequacy of decisions related to CBM.

### **Considering Alternative PM Schedules: Jenkins et al. (2017)**

Jenkins and colleagues (2017) hypothesized that intermittent PM schedules could decrease time commitments related to assessment administration and, therefore, enhance the feasibility of using CBM and the DBI process for teachers. Jenkins et al. argued that decreasing the time demands related to CBM administration would provide more time for teachers to use the data to inform instruction, particularly for students with inadequate response to intervention. However, the authors also realized the importance of identifying PM schedules that are both sufficiently accurate (i.e., identify the adequacy of student growth) and timely (i.e., in the fewest weeks). Therefore, Jenkins et al. investigated different PM schedules for CBM in reading for students with disabilities to identify schedules that may decrease time commitments for teachers while also maintaining sufficiently high DBI accuracy and timeliness necessary for the DBI process.

Jenkins et al. (2017) administered three ORF probes each week to 56 students (demographics in Table 1). From the total set of 42 passages administered to each student, Jenkins et al. simulated six PM schedules (i.e., one a week, two every 2 weeks, and three every 3, 4, 5, and 6 weeks). They estimated students' TG slope by inputting all 42 scores. They also estimated the slopes of the data for each PM schedule across weeks, using relevant CBM probes up to the week to date (see our Method section for more detail). Jenkins et al. called the slopes from the PM schedules "weekly slopes" because the data were collected on one day each week.

Jenkins et al. (2017) established a goal growth rate of 1.0-WRC increase each week, citing this goal as a reasonable rate of growth for second to sixth graders (Deno et al., 2001). For each week, Jenkins et al. assessed whether TG and weekly slopes indicated adequate or inadequate progress compared to the goal of 1.0-WRC increase per week. They determined each PM schedule's decision accuracy across weeks by calculating the percentage of the sample for whom the TG and weekly slope indicated the same student response (i.e., adequate or inadequate). Jenkins et al. also reported the number of weeks it took PM schedules to reach 70% and 75% accuracy as potential thresholds of sufficiently high decision accuracy.

Overall, Jenkins et al. (2017) reported that PM schedules had similar levels of decision accuracy and that more intermittent PM schedules did not undermine the timeliness of instructional decision making. The authors interpreted their results as demonstrating that intermittent PM schedules could potentially be used instead of the traditional, weekly PM schedule. They suggested that presenting these intermittent schedules as options to teachers could address the primary teacher-reported barrier of lack of time to engage in data-based decision making.

### **Replication in Special Education Research**

Replication is an important component of the empirical process that generates scientific findings and contributes to the understanding

of broader theories (Coyne et al., 2016). However, replication studies are underrepresented in the literature base, constituting only 0.41% of articles in special education journals (Lemons et al., 2016). Coyne et al. (2016) classify replications as either direct or conceptual. Direct replications are studies in which the researchers replicate all aspects of the original study (e.g., all aspects of the methods and analyses). Direct replication is challenging to conduct in special education, particularly because of the nature of applied school settings and differences across participant samples. In contrast, conceptual replications are studies in which the researchers replicate aspects of the original study but accept variability in some aspects (e.g., altering the participant sample).

To the best of our knowledge, there are limited studies related to CBM in special education that have been explicitly described as replication studies (e.g., Conoyer et al., 2019; J. Hosp et al., 2018). Some researchers have conducted CBM-related research studies and have discussed the extent to which the results replicate or extend findings from previous studies with similar aims (e.g., Diggs & Christ, 2019; January & Ardoin, 2015). However, in these instances, the authors did not frame their investigations as a specific replication study.

In contrast, both Conoyer et al. (2019) and J. Hosp et al. (2018) designed conceptual replication studies that assessed the technical adequacy of CBM tools with different samples of students than the respective original studies. In line with best practices for replication, the authors of both studies explicitly discussed the alignment of research procedures with the original procedures and explicitly compared results of the original and replication study. Given the importance of replication to the empirical process and the wide use of CBM for data-based decision making in schools, there is a need for more CBM-related replication research studies.

### **Purpose**

The purpose of this study was to replicate and extend the work of Jenkins et al. (2017), the findings of which could have broad implications for schools. We attempted to replicate

**Table 1.** Student Demographics.

Variable	Current study				Jenkins et al. (2017)			
	<i>M</i>	<i>SD</i>	<i>n</i>	%	<i>M</i>	<i>SD</i>	<i>n</i>	%
Age	9.45	0.88			NR	NR		
Grade	3.25	0.80			4.23	0.95		
Second			11	21.57			1	1.79
Third			16	31.37			11	19.64
Fourth			24	47.06			24	42.86
Fifth			—	—			14	25.00
Sixth			—	—			6	10.71
Instructional reading level	1.90	0.77			2.80	0.90		
First grade			13	25.49			2	1.79
Second grade			28	54.90			21	19.64
Third grade			8	15.69			21	42.86
Fourth grade			2	3.92			10	25.00
Fifth grade			—	—			2	10.71
Gender								
Female			14	27.45			20	35.71
Ethnicity ( <i>N</i> = 46)								
Hispanic			20	43.48			NR	NR
Race ( <i>N</i> = 50)								
White			25	50.00			NR	NR
Black			23	46.00			NR	NR
Hispanic (teacher 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		
Final (Week 13)	61.51	33.39			NR	NR		

Note. Total *N* = 51 unless noted for current study. Original study had a final sample of 56. EBD = emotional and behavioral disorders; EL = English learner; FDD = functional or developmental delay; IDD = intellectual or developmental disability; IEP = individualized education program; LD = learning disability; SLI = speech or language impairment; NR = not reported; OHI = other health impairment; SEL = social and emotional learning; WRC = words read correctly.

Jenkins et al.'s procedures as closely as possible; however, there were critical differences (see Table 2). Therefore, the current study can best be categorized as a closely aligned conceptual replication of Jenkins et al.'s study.

*Therefore, the current study can best be categorized as a closely aligned conceptual replication of Jenkins et al.'s study.*

In October 2018, we preregistered our research plan (e.g., intended sample, data collection procedures, and data analysis) with the Open Science Foundation. An anonymous version of the preregistration is publicly available at <https://bit.ly/2O8KkKX>. Throughout the timeline of our study, we made changes from the preregistered plan. We outline these changes in Table 3 and provide a rationale for each change. To enhance readability, we report our method and results in alignment with the changes that we made.

We considered Jenkins et al.'s (2017) research questions as the primary research questions for our direct replication. These included "Is decision-making accuracy from intermittent PM inferior to that from weekly PM, the current standard?" and "How many weeks of PM do these schedules require to reach specific levels of decision accuracy?" (Jenkins et al., 2017, pp. 45, 47). We acknowledge that we could have omitted the term "inferior" from the first research question to highlight the goal of comparing the PM schedules without implying a presumed direction; however, in the spirit of replication, we opted to keep Jenkins et al.'s original wording.

We extended the first research question by considering a priori accuracy thresholds (70% and 75% accuracy). We selected these thresholds based on the two accuracy thresholds explored by Jenkins et al. (2017). Although these thresholds may be low for special education eligibility decisions (which have higher stakes related to resource allocation), they are more acceptable thresholds for decisions related to students' progress toward annual goals and objectives in that they strike a reasonable balance between accurate yet timely instructional decisions. In follow-up analyses,

we contrasted PM schedules only when at least one schedule met the required threshold.

We extended the second research question by considering whether the time it took intermittent PM schedules to reach each accuracy threshold was within 2 weeks of the time it took weekly PM schedules to reach the same threshold. We selected the 2-week criterion as we hypothesized that instructional changes made at any point within this brief window of time would not lead to differences in student outcomes. Overall, we assessed the need to account for teacher-reported instructional changes in our analyses and compared our results to the original results to assess whether the findings and interpretations of findings were the same.

## Method

### Sample

After obtaining institutional review board approval, we recruited 12 special educators from six elementary schools within an urban district in the southeastern United States. All teachers worked predominantly with students identified with high-incidence disabilities. The majority of the teachers were White ( $n = 9$ ; 75%) and female ( $n = 11$ ; 91.7%). These teachers helped recruit 64 students for this study. At the conclusion of the study, we provided teachers their students' data.

Eight students moved prior to the end of data collection. Following Jenkins et al.'s (2017) data-cleaning procedures, we excluded students missing more than 1 week of data from the final data analyses. The final sample included 51 students (14 female; 27.45%), nine (17.65%) of whom missed 1 week of data collection. Results from a  $t$  test indicated that the mean difference between TG for students with incomplete versus complete data (0.61 and 0.88, respectively) was not statistically significant,  $t(49) = -1.38, p = .17$ . The effect size difference between these groups, however, was  $d = -0.28$ , suggesting that students with missing data, on average, demonstrated poorer growth than students with complete data. See Table 1 for demographics for this sample and the reported data for the original study's sample.

**Table 2. Differences Between Jenkins et al. (2017) and the Current Study.**

Area	Jenkins et al. (2017)	Current study	Rationale for change
Subject characteristics	<p>Recruited: 66; final: 56                      Attrition: 15.15%                      Missed 1 week: 8.93%                      Mean grade: 4.23                      Mean reading level: 2.8                      Disabilities: See Table 1                      Mean TG rate: 1.12 (SD = 0.88)                      Nonresponder rate: 45%</p> <p>Jenkins and colleagues reported their teachers were not using CBM</p>	<p>Recruited: 64; final: 51                      Attrition: 20.31%                      Missed 1 week: 17.65%                      Mean grade: 3.25                      Mean reading level: 1.9                      Disabilities: See Table 1                      Mean TG rate: 0.84 (SD = 0.55)                      Nonresponder rate: 68.63%</p> <p>Districtwide policies required teachers to collect PM assessment data for all students identified for special education. Solution: Gave teacher surveys and additional analyses</p>	<p>Differences in the setting and the population of students within each setting</p>
School-based CBM	<p>Jenkins and colleagues reported their teachers were not using CBM</p>	<p>Districtwide policies required teachers to collect PM assessment data for all students identified for special education. Solution: Gave teacher surveys and additional analyses</p>	<p>Teacher completed surveys of instructional changes across study duration. We conducted point biserial correlations to determine if student growth slopes correlated significantly with reported changes (thereby indicating a need to control for those changes). Correlation was not significant.</p>
Passage selection	<p>Supplemented AIMSweb passages with Edcheckup passages (www.edcheckup.com)</p> <p>Instructional level: Determined by teacher estimates only</p>	<p>Readministered randomly ordered AIMSweb passages</p> <p>Instructional level: Determined by teacher estimate, followed checks that baseline scores between 10th and 50th percentiles for grade level of passages</p>	<p>Passages assigned the same grade-level difficulty across CBM vendors are not necessarily functionally equivalent. Majority of students had low potential for practice effect given 10 weeks from initial to repeat read</p>
Data collection	<p>Fall data collection (no reported school break in the middle of the study weeks)</p> <p>CBM administration</p> <p>Baseline: 3 probes                      Weeks 1-11: 3 probes                      Week 12: 6 probes</p> <p>No reported makeup day if students were absent</p>	<p>January to April data collection (spring break week in the middle of the study weeks)</p> <p>CBM administration</p> <p>Baseline: 3 probes                      Weeks 1-13: 3 probes                      Makeup day if students were absent on assigned testing day</p>	<p>Add data to support teacher estimates of instructional level and to ensure students read passages sensitive to growth</p> <p>Recruitment time and district schedule</p> <p>School requirements limited time to assess within and across schools, making six assessments per student in Week 12 impossible</p> <p>Minimize missing data</p>
Interobserver agreement	<p>Minimum: 4 students per examiner in Weeks 2 and 3 and Weeks 9 and 10</p> <p>1 every week, 2 every 2 weeks, 3 every 3, 4, 5, and 6 weeks</p>	<p>Audio of 30% of all passages blindly and independently double-scored</p> <p>1 every week, 2 every 2 weeks, 3 every 3, 4, 5, and 6 weeks, plus three additional: 1 every week (second probe)                      1 every week (third probe)                      1 every 2 weeks</p>	<p>Limited resources to allow for live reliability observations</p> <p>Increased percentage of passages for increased confidence in scoring</p> <p>Compare the weekly PM schedule accuracy to other hypothetical versions of weekly schedules to determine the potential presence of passage order effect</p>
PM schedule	<p>1 every week, 2 every 2 weeks, 3 every 3, 4, 5, and 6 weeks</p>	<p>1 every week (second probe)                      1 every week (third probe)                      1 every 2 weeks</p>	<p>Compare an alternative, common intermittent PM schedule employed by local schools</p>
Results	<p>Reported Week 4+ results</p> <p>Higher decision accuracy, greater significance of binomial tests, less time to thresholds</p>	<p>Reported all results</p> <p>Lower decision accuracy, greater significance of binomial tests, greater time to thresholds</p>	<p>More complete reporting of data across time</p> <p>Differences in the results may be due to sample differences</p>

Note. CBM = curriculum-based measurement; PM = progress monitoring; TG = true growth.

**Table 3.** Differences Between Preregistration and the Final Study.

Area	Preregistration	As conducted	Rationale for change
Data collection timeline	Fall	January to April	Recruitment took longer than anticipated
Recruitment sample size	100	64	Due to data collection timeline, insufficient weeks remaining in school year to meet recruiting goals
CBM probes per week	Baseline: 3 probes Weeks 1–11: 3 probes Week 12: 6 probes	Baseline: 3 probes Weeks 1–13: 3 probes	School requirements limited time to assess within and across schools, precluding us from conducting six assessments per student in Week 12
Reliability observations	At least four students per examiner per observation window (Weeks 2 and 3 and Weeks 9 and 10)	30% of all passages independently and blindly double-scored via audio recordings	Due to limited personnel, we decided to evaluate reliability via audio; due to this change, we were compelled to account for greater proportion of reliability observations than initially planned
Teacher surveys	Weekly surveys of instructional changes	Midpoint and final survey only	Teachers reported not being able to complete the surveys on a weekly basis
Determining need to control for instructional changes	Logit regression to determine correlation between teacher-reported instructional changes and student growth slopes	Point biserial correlations	Logit regression is best applied to data sets in which the outcome—not predictor—is categorical
Statistical test of differences in PM schedules' time to accuracy threshold	Repeated-measure ANOVA (within-subjects factor: PM schedule; dependent variable: time to accuracy threshold)	Did not conduct ANOVA	Due to the nature of the data set (i.e., time to accuracy threshold variable was a single value for each PM schedule), this analysis could not be conducted

Note: CBM = curriculum-based measurement; PM = progress monitoring.



## Materials

Members of our research team administered a total of 42 PM probes (AIMSweb; Shinn et al., n.d) to each student. We used a random-number generator to assign each student a random sequence of passages. Because there were only 33 passages per grade level (23 passages for first grade), we readministered the randomly ordered passages after all instructional-level passages had been administered. We elected to readminister passages, rather than supplement with another vendor's passages, because passages assigned the same grade-level difficulty across CBM vendors are not necessarily functionally equivalent (Ardoin & Christ, 2009; Ford et al., 2017).

Readministering passages increases the potential for practice effects (Jenkins et al., 2005). However, the evidence suggests that practice effects are negligible after a 10-week interval between initial and follow-up administration (Jenkins et al., 2005). The majority of our sample did not begin repeating passages until 12 weeks had passed. For the 25.49% of students reading at a first-grade instructional level, repeated reading of passages began in the third passage of the 8th week. Results from a *t* test indicated that the mean difference between TG for students reading at a second- to fourth-grade level versus a first-grade level was not statistically significant,  $t(49) = 1.60$ ,  $p = .12$ . The difference between TG for students reading at a second- to fourth- versus a first-grade level was  $d = 0.28$ , indicating that students reading at a first-grade instructional level, on average, demonstrated poorer growth than students reading at higher levels. These data suggest minimal risk of practice effect.

## Procedures

Three graduate student research assistants (RAs; two female) served as examiners for this study. We 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, and an administration checkout. All RAs obtained 98% accuracy or greater on scoring WRC.

Data collection began in the second week of January. During the first week, RAs determined students' instructional reading level by administering passages at the teacher-recommended instructional level. If the student's median score for these three passages was not between the 10th and 50th percentiles, the RA administered passages at the next lower or higher grade level until reaching the 10th-to-50th-percentile criterion. RAs followed this procedure to place students in an instructional level that was appropriate for PM. Note that we did not use a traditional definition of instructional level (i.e., 93% to 97% accuracy; Gickling & Armstrong, 1978; Treptow et al., 2007). Instead, our procedures were in line with Jenkin et al.'s (2017) rationale, which emphasized the need to ensure sensitivity to growth in PM data (Filderman & Toste, 2017).

The data from the passages corresponding to the students' identified instructional level were used as the student's first three data points. In subsequent weeks, RAs administered the remaining AIMSweb passages at each student's instructional level. After baseline week, RAs administered three randomly ordered CBM passages a week to each student for 13 additional weeks of data collection from January to April. We assigned each student a consistent testing day in the middle of each week. If a student was absent, RAs returned on Friday for makeup assessments (Mondays during weeks without school on Friday). There was a 1-week break from data collection (between Weeks 9 and 10) during the district's spring break.

RAs audio recorded each test administration. Students read a student version of the passage, and RAs recorded student responses on the examiner version, which included a word count along the margins. Consistent with the original study's procedures, RAs told participants the following:

It's time for a short reading check. I'm using a timer to remind me how long we need to listen. When we 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, and RAs recorded errors (i.e., mispronunciations, skipped words, and hesitations >3s). In the case of hesitations, RAs provided students with the word after 3 s. 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 administering the three passages for the week, RAs thanked the student and returned them to their classroom. 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.

A second scorer used the paper records and rescored each of these passages for interscorer reliability of WRC scores. Interscorer reliability was high (96.08%). RAs also blindly and independently double-scored a random sample of 30% of probes from audio to assess interobserver reliability. Following Jenkins et al.'s (2017) protocol, we calculated interobserver reliability by dividing the lower by the higher WRC score from the lead and reliability data. We averaged these values across all reliability passages. Interobserver agreement was high (97.18%).

Participating schools were also conducting PM of students with a variety of assessments. To account for instructional changes that teachers may have made in relation to these data, teachers completed a survey of reading instruction three times across the study. On this survey, teachers provided information about students' reading instruction (e.g., session length and frequency, grouping type, and time dedicated to each area of reading instruction).

### *Design and Analysis*

*Replicated Analyses.* We conducted the same primary data analyses as those employed by Jenkins and colleagues (2017). Prior to

conducting the analyses, we cleaned the data by excluding data of participants who missed more than 1 week of data collection, replicating the original study's procedures. We conducted all analyses in Stata/SE 14.0 (StataCorp, 2015).

*Growth estimates and PM schedules.* We conducted an ordinary least squares (OLS) regression using all 42 CBM data points to obtain a TG estimate for each student. To account for each data point, we followed Jenkins and colleagues' (2017) procedure of using individual scores in all slope calculations, adding 0.003 days (5 min) to the day-of-administration variable for each additional measure administered in the same day.

We also calculated weekly slopes for all PM schedules by running the OLS regression with the available data that (a) had been collected up to that point in time and (b) fit the respective PM schedule. We included the three baseline probes 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). We examined the same intermittent PM schedules analyzed by Jenkins and colleagues (2017), which included (a) one CBM weekly, using the first passage administered each week; (b) two CBMs every 2 weeks, using the first two passages administered in Weeks 2, 4, 6, 8, 10, and 12; (c) three CBMs every 3 weeks, using all CBMs administered in Weeks 3, 6, 9, and 12; (d) three CBMs every 4 weeks, using all CBMs administered in Weeks 4, 8, and 12; (e) three CBMs every 5 weeks, using all CBMs administered in Weeks 5 and 10; and (f) three CBMs every 6 weeks, using all CBMs administered in Weeks 6 and 12.

*Assessing adequacy of student growth.* After conducting all OLS regressions, we assessed the adequacy of student growth as determined by TG (the slope that took into account all 42 CBM probes). If a student's TG slope met or exceeded the goal of 1.0 WRC per week, that student would have been designated as demonstrating adequate growth. If the

student's TG slope was less than 1.0-WRC increase per week, that student would have been designated as demonstrating inadequate growth. We created a dichotomized "adequate growth" variable to indicate the adequacy of each student's TG (1 = adequate growth; 0 = inadequate growth). Next, we assessed the adequacy of each student's growth as determined by the weekly slopes from the relevant PM schedules each week. We created a dichotomized adequate-growth variable for each of these weekly slopes as well.

**Decision accuracy.** We compared the dichotomized adequate-growth variable for each PM schedule's weekly slope with the dichotomized adequate-growth variable for TG and determined whether the values matched or not. Matched decisions indicated that both the weekly slope and TG determined adequate or inadequate growth. Mismatched decisions occurred when the PM schedule's weekly slope indicated adequate growth but TG indicated inadequate growth (a missed nonresponder) or vice versa (a missed responder). We created a "decision match" variable. Finally, we determined decision accuracy by calculating the proportion of matched decisions for each PM schedule across students.

#### *Other Jenkins et al. (2017) data analyses.*

Following Jenkins et al.'s (2017) procedure, we ran a binomial test for each schedule's decision accuracy by week to calculate whether obtaining each accuracy level or higher was significantly above chance (i.e., 50%). We also calculated the correlation between (a) TG slopes and student grade level and (b) TG slopes and student instructional level. We ran descriptive statistics to report the sample's average TG slopes, the standard deviation of those slopes, and the skewness of the distribution of the TG slopes across participants. Last, we calculated the number of participants failing to achieve the TG goal rate of 1.0-WRC increase or greater per week across study weeks.

**Supplemental Data Analyses.** We extended Jenkins et al.'s (2017) analyses by calculating

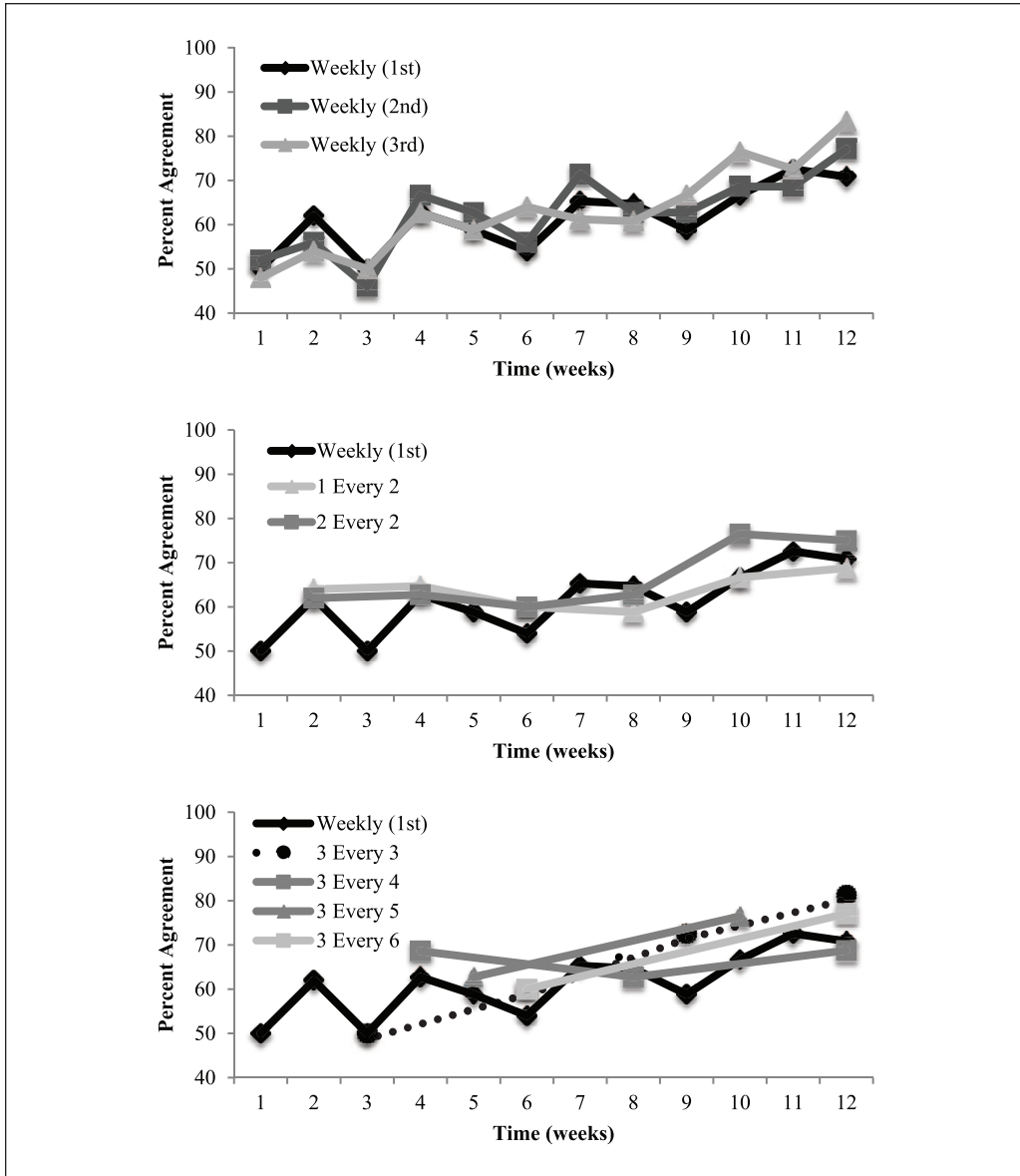
weekly slopes and assessing the accuracy of three additional PM schedules. These included alternative versions of the weekly PM schedule (i.e., using the second or third CBM administered each week) and one CBM every 2 weeks (using the first passage administered in Weeks 2, 4, 6, 8, 10, and 12).

We also conducted two supplemental analyses that extended the work of Jenkins and colleagues (2017). First, we ran point biserial correlations to determine whether there was a significant correlation between teacher-reported instructional changes and students' TG slope. We 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, we calculated two point biserial correlations: (a) between students' TG slope and teacher-reported instructional changes between the midpoint and final survey and (b) between students' TG 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 TG slope, we planned to control for teacher-reported instructional changes in the primary OLS regression analyses.

## Results

Over the 14-week period, the sample's mean TG was 0.84 words per week ( $SD = 0.55$ ). The distribution of TG slopes across participants was approximately symmetrical, with a nonsignificant skewness of 0.22. Although the majority (68.63%) of the current sample failed to achieve the goal growth rate, TG was not significantly correlated with grade ( $r = .04$ ) or instructional reading level ( $r = .17$ ).

Teachers reported providing reading intervention to students 5 days a week (session length  $M = 45.34$  min,  $SD = 15.34$  min) in small groups ( $M = 4.14$  students). Additionally, teachers reported that their instruction focused primarily on reading 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 reading instruction (intervention context,



**Figure 1.** Decision accuracy across progress-monitoring (PM) schedules. The data for the weekly (first probe) PM schedule are on each graph for ease of comparison with other PM schedules.

grouping, or reading emphasis) for 21 students (41.18%) on at least one of their completed surveys. The results of the point biserial correlation tests indicated that there was not a significant relation between students' TG and teacher-reported instructional changes between Weeks 1-6, between Weeks 7-12, or across the entire study duration. Therefore, we

did not control for instructional changes in our analyses.

### Decision Accuracy

Figure 1 shows the decision accuracy of PM schedules across the weeks of the study. The accuracy of the traditional weekly PM

schedule (weekly [first probe]) is represented in each of the graphs as a comparison for each of the alternative PM schedules analyzed. For all PM schedules, decision accuracy increased across time, though imperfectly due to variability of accuracy across weeks for each PM schedule. Table 4 shows the accuracy of PM schedules across the weeks of the study. Each week, we sorted accuracy from most to least accurate and shaded the traditional, weekly PM schedule row gray. Table 4 also shows the percentage overlap between the 39 TG passages administered after baseline and the number of passages contributing to each PM schedule's weekly slope calculations. Finally, Table 4 reports the results of the binomial tests.

Jenkins et al. (2017) defined a contrast as each comparison between the weekly PM schedule and any intermittent PM schedule assessed in the same week. The results indicated that intermittent PM schedules were at least as accurate as the weekly PM schedule in 12 of the 15 contrasts from Week 4 on (the weeks Jenkins et al. also reported). 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 at least one of the contrasted schedules reached the 70% accuracy threshold, the intermittent schedules were more accurate than the weekly schedule. In five contrasts, at least one of the schedules reached the 75% accuracy threshold. In all of these contrasts, the intermittent PM schedule was more accurate than the weekly schedule.

We descriptively examined the types of errors in PM schedules' data. We tracked the instances in which a PM schedule misidentified a student whose TG data showed inadequate growth (missed nonresponder) and the instances in which a PM schedule misidentified a student whose TG data showed adequate response (missed responder). Of the mismatched decisions, 63.39% were instances in which a PM schedule missed a nonresponder. For nearly every week and PM schedule, there was a higher prevalence of missed nonresponders versus missed responders. A figure with these data is available online in supplemental materials.

We ran the analyses on three additional PM schedules (i.e., every week [second probe], every week [third probe], and one every 2 weeks) not explored by Jenkins and colleagues (2017). 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 demonstrated superior accuracy compared to one another in four of the 12 weeks. Additionally, the one-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 3 of the 6 applicable weeks, equally accurate in 1 week, and less accurate in 2 weeks.

### *Timeliness*

Table 5 shows the number of weeks it took each PM schedule to reach 70% and 75% accuracy the first time and compares those results to the original study's results. The every-3-weeks PM schedule reached 70% accuracy the earliest (Week 9) and reached 75% the next time that schedule was assessed (Week 12). The majority of the PM schedules reached 70% and 75% accuracy between Weeks 10 and 12, though the weekly (first probe) PM schedule never reached 75% accuracy. Neither the every-4-weeks nor the one-every-2-weeks schedule reached either accuracy threshold.

### **Discussion**

The purpose of this study was to replicate and extend the work of Jenkins and colleagues (2017). We explored the relative accuracy and timeliness of PM schedules with a sample of 51 second to fourth graders receiving special education services. Overall, results demonstrated that intermittent PM schedules had greater accuracy and better timeliness compared with weekly PM schedules in almost all incidences, replicating the conclusions asserted by Jenkins et al.

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

Decision point: PM schedule	Accuracy (%)	Jenkins et al. (2017) accuracy	Score overlap (%)	Current sample slope, <i>M</i> ( <i>SD</i> )
<i>True growth</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>0.84 (0.55)</i>
Deciding at Week 1				
Weekly ( <i>second probe</i> )	52.0	<i>N/A</i>	2.6	<i>1.08 (11.37)</i>
Weekly ( <i>first probe</i> )	50.0	NR	2.6	<i>-0.78 (11.23)</i>
Weekly ( <i>third probe</i> )	48.0	<i>N/A</i>	2.6	<i>5.77 (15.29)</i>
Deciding at Week 2				
<i>1 every 2 weeks</i>	64.0*	<i>N/A</i>	2.6	<i>0.44 (5.74)</i>
Weekly ( <i>first probe</i> )	62.0	NR	5.1	<i>0.30 (5.47)</i>
<i>2 every 2 weeks</i>	62.0	NR	5.1	<i>1.02 (4.36)</i>
Weekly ( <i>second probe</i> )	56.0	<i>N/A</i>	5.1	<i>1.53 (4.75)</i>
Weekly ( <i>third probe</i> )	54.0	<i>N/A</i>	5.1	<i>0.93 (4.85)</i>
Deciding at Week 3				
Weekly ( <i>first probe</i> )	50.0	NR	7.7	<i>1.61 (3.60)</i>
Weekly ( <i>third probe</i> )	50.0	<i>N/A</i>	7.7	<i>0.46 (3.59)</i>
<i>3 every 3 weeks</i>	50.0	NR	7.7	<i>0.94 (3.30)</i>
Weekly ( <i>second probe</i> )	46.0	<i>N/A</i>	7.7	<i>0.63 (4.48)</i>
Deciding at Week 4				
<i>3 every 4 weeks</i>	68.6**	71.4**	7.7	<i>0.99 (2.88)</i>
Weekly ( <i>second probe</i> )	66.7*	<i>N/A</i>	10.3	<i>0.86 (3.42)</i>
<i>1 every 2 weeks</i>	64.7*	<i>N/A</i>	5.1	<i>0.69 (4.31)</i>
Weekly ( <i>first probe</i> )	62.7*	64.3*	10.3	<i>1.15 (3.18)</i>
Weekly ( <i>third probe</i> )	62.7*	<i>N/A</i>	10.3	<i>0.79 (2.73)</i>
<i>2 every 2 weeks</i>	62.7*	66.1*	10.3	<i>0.89 (3.30)</i>
Deciding at Week 5				
Weekly ( <i>second probe</i> )	62.7*	<i>N/A</i>	12.8	<i>0.96 (3.11)</i>
<i>3 every 5 weeks</i>	62.7*	71.4**	7.7	<i>0.91 (2.36)</i>
Weekly ( <i>first probe</i> )	58.8	58.9	12.8	<i>0.86 (2.54)</i>
Weekly ( <i>third probe</i> )	58.8	<i>N/A</i>	12.8	<i>0.88 (2.65)</i>
Deciding at Week 6				
Weekly ( <i>third probe</i> )	64.0*	<i>N/A</i>	15.4	<i>1.03 (2.33)</i>
<i>2 every 2 weeks</i>	60.0	73.2**	15.4	<i>0.38 (2.29)</i>
<i>1 every 2 weeks</i>	60.0	<i>N/A</i>	7.7	<i>0.41 (2.90)</i>
<i>3 every 3 weeks</i>	60.0	76.8**	15.4	<i>0.62 (2.10)</i>
<i>3 every 6 weeks</i>	60.0	78.7**	7.7	<i>0.62 (2.11)</i>
Weekly ( <i>second probe</i> )	56.0	<i>N/A</i>	15.4	<i>0.62 (2.35)</i>
Weekly ( <i>first probe</i> )	54.0	66.1*	15.4	<i>0.57 (2.23)</i>
Deciding at Week 7				
Weekly ( <i>second probe</i> )	71.4**	<i>N/A</i>	17.9	<i>0.81 (1.98)</i>
Weekly ( <i>first probe</i> )	65.3*	NR	17.9	<i>0.78 (1.79)</i>
Weekly ( <i>third probe</i> )	61.2	<i>N/A</i>	17.9	<i>0.94 (2.36)</i>
Deciding at Week 8				
Weekly ( <i>first probe</i> )	64.7*	71.4**	20.5	<i>0.56 (1.51)</i>
Weekly ( <i>second probe</i> )	62.7*	<i>N/A</i>	20.5	<i>0.84 (1.75)</i>
<i>2 every 2 weeks</i>	62.7*	73.2**	20.5	<i>0.56 (1.61)</i>
<i>3 every 4 weeks</i>	62.7*	67.9*	15.4	<i>0.82 (1.50)</i>
Weekly ( <i>third probe</i> )	60.8	<i>N/A</i>	20.5	<i>0.94 (1.95)</i>
<i>1 every 2 weeks</i>	58.8	<i>N/A</i>	10.3	<i>0.39 (1.93)</i>

(continued)

**Table 4. (continued)**

Decision point: PM schedule	Accuracy (%)	Jenkins et al. (2017) accuracy	Score overlap (%)	Current sample slope, <i>M</i> ( <i>SD</i> )
<i>True growth</i>	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>	<i>0.84 (0.55)</i>
Deciding at Week 9				
3 every 3 weeks	72.5**	76.8**	23.1	0.80 (1.11)
<i>Weekly (third probe)</i>	66.7*	<i>N/A</i>	23.1	0.98 (1.39)
<i>Weekly (second probe)</i>	62.7*	<i>N/A</i>	23.1	0.92 (1.24)
<i>Weekly (first probe)</i>	58.8	66.1*	23.1	0.49 (1.29)
Deciding at Week 10				
<i>Weekly (third probe)</i>	76.5**	<i>N/A</i>	25.6	0.84 (1.09)
2 every 2 weeks	76.5**	76.8**	25.6	0.63 (0.94)
3 every 5 weeks	76.5**	73.2**	15.4	0.67 (0.89)
<i>Weekly (second probe)</i>	68.6**	<i>N/A</i>	25.6	0.90 (0.93)
<i>Weekly (first probe)</i>	66.7*	75*	25.6	0.49 (1.12)
1 every 2 weeks	66.7*	<i>N/A</i>	12.8	0.46 (1.36)
Deciding at Week 11				
<i>Weekly (first probe)</i>	72.5**	NR	28.2	0.55 (0.93)
<i>Weekly (third probe)</i>	72.5**	<i>N/A</i>	28.2	0.91 (0.89)
<i>Weekly (second probe)</i>	68.6**	<i>N/A</i>	28.2	0.92 (0.84)
Deciding at Week 12				
<i>Weekly (third probe)</i>	83.3**	<i>N/A</i>	30.8	0.96 (0.82)
3 every 3 weeks	81.3**	89.3**	30.8	1.01 (0.90)
<i>Weekly (second probe)</i>	77.1**	<i>N/A</i>	30.8	1.03 (0.82)
3 every 6 weeks	77.1**	83.9**	15.4	1.09 (1.05)
2 every 2 weeks	75.0**	83.9**	30.8	0.89 (0.87)
<i>Weekly (first probe)</i>	70.8**	78.6**	30.8	0.66 (0.89)
1 every 2 weeks	68.8**	<i>N/A</i>	15.4	0.68 (1.15)
3 every 4 weeks	68.8**	83.9**	23.1	1.00 (2.96)

Note. Shaded area indicates the results of the traditional, weekly curriculum-based measurement 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. Time to Accuracy Thresholds.**

PM schedule	Time to accuracy threshold in weeks			
	70% accuracy		75% accuracy	
	Current	Jenkins et al. (2017)	Current	Jenkins et al. (2017)
Weekly (first probe)	11	8	Never	10
Weekly (second probe)	12	<i>N/A</i>	12	<i>N/A</i>
Weekly (third probe)	10	<i>N/A</i>	10	<i>N/A</i>
2 every 2 weeks	10	6	10	10
1 every 2 weeks	Never	<i>N/A</i>	Never	<i>N/A</i>
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



### *Does Intermittent PM Undermine Decision Accuracy?*

Of the schedules also evaluated by Jenkins et al. (2017), there was 0% to 14.5% difference between the most and least accurate schedule in a week. Intermittent PM schedules were at least as accurate as the traditional, weekly PM schedule in the majority of weeks and in the majority of weekly versus intermittent contrasts. Further, as Jenkins et al. reported, the every-3-weeks PM schedule either tied for or was the most accurate schedule across all relevant weeks (see Table 4). These results are in line with our initial hypothesis that decision-making accuracy from intermittent PM would be indeterminately different from that of weekly PM. These results provide preliminary evidence for the comparability of intermittent and weekly PM schedules. Note, we used all 42 CBM probes to calculate each student's TG slope. We calculated each student's weekly slope by using the relevant subset of probes from the same pool of 42 probes. Thus, the overlap between data used to calculate TG and weekly slopes increased across time (see Table 4). This overlap contributed, in part, to the improved decision accuracy of every PM schedule over time.

*Intermittent PM schedules were at least as accurate as the traditional, weekly PM schedule in the majority of weeks*

We extended the first research question by considering whether the comparability of weekly versus intermittent PM schedules differed when considering only comparisons of schedules in which at least one schedule met an a priori accuracy threshold of 70% or 75%. Fewer than half of the weekly-versus-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. Although these descriptive results do not originate from a statistical test comparing PM schedules, it provides preliminary evidence that counters our 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 the nature of the PM schedules themselves, given that 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 and testing context (e.g., variability in CBM passages and contextual differences between sessions) compared to PM schedules that aggregate multiple data points within a week (see Yoder et al., 2018, p. 56). It is possible that this aggregation effect factors into the finding in both the current and original study that the every-3-weeks PM schedule—which accounted for the same number of passages as the weekly schedule—was consistently more accurate than the weekly schedule.

### *Do Students Perform More Poorly on Initial Passages Administered?*

Jenkins et al. (2017) 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. We examined alternative weekly PM schedules using the second and third probe given each week. The three versions of the weekly PM schedule had comparable accuracy (see Figure 1). These results suggest that the poorer accuracy of the weekly versus intermittent PM schedules may relate to passage variability, the effect of which can be attenuated by aggregating data at a given time point. This hypothesis aligns with the principle of aggregation in classical measurement theory in 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 9 to 12 weeks to reach the 70% and 75% accuracy thresholds explored by Jenkins and colleagues (2017). This length of time was 2 to 3 weeks longer than the amount of time we 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. The results of the current study are in line with previous research suggesting that longer durations of PM reduced the standard error of slopes and corresponding confidence intervals in CBM, particularly in assessment contexts that were less optimally controlled, such as classrooms with other activities happening while test administration is occurring (Christ, 2006).

Our results, however, should be couched in a broader consideration of whether these thresholds are the most appropriate or desirable thresholds to consider. For the purpose of replication, we used the same accuracy thresholds that Jenkins et al. (2017) used. 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 reflect students’ true performance.

We extended the second research question by considering whether the time it took intermittent PM schedules to reach each accuracy threshold was within 2 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 2 weeks of the time it took the weekly PM schedule to reach the same accuracy threshold in nearly

every instance (see Table 5). The time it took intermittent PM schedules to reach the 75% accuracy threshold was also less than or within 2 weeks of the time it took the weekly PM schedule to reach the higher accuracy threshold in all instances where this comparison was possible to assess (see Table 5). It was not possible to make all higher-accuracy-threshold comparisons, however, because the weekly PM schedule (and some intermittent PM schedules) never reached 75% accuracy.

### *Do the Results of This Study Replicate the Original Findings?*

Jenkins and colleagues (2017) 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. Jenkins et al. also found that it took intermittent PM schedules 4 to 6 weeks to reach 70% accuracy and 6 to 12 weeks for all intermittent PM schedules (except the every-5-weeks schedule) to reach 75% accuracy. Jenkins et al. found that the weekly PM schedule took 8 and 10 weeks to reach 70% and 75% accuracy, respectively. In this replication study, PM schedules took longer than reported by Jenkins et al. to reach accuracy thresholds (by more than 2 weeks) in nearly all instances (see Table 5). Despite this difference, 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 findings also do not appear to be influenced by teacher-reported instructional changes, as the results of the point biserial correlation tests indicated no need to account for these changes in our analyses.

### *How Does This Study Compare to Jenkins et al. (2017)?*

There were a few differences between the current study and Jenkins et al.’s (2017) study (see Table 2). First, there were dissimilarities in the sample that are important to note. Despite our recruitment efforts, our final sample was slightly smaller than Jenkins et al.’s

final sample (51 vs. 56 students), though we recruited a similar number of students initially (64 vs. 66). The sample of students recruited for this study consisted of students from transient families with histories of frequent moves, students who demonstrated chronic absenteeism, and students who experienced instability in home life (e.g., placement into foster care). As a result, there was a higher attrition rate in this study than in Jenkins and colleagues' study (20.31% vs. 15.15% attrition). 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%). Although the *t* test indicated that the difference in TG for students with incomplete versus complete data was not significant, there was an effect size difference of  $d = -0.28$  between the groups. With a larger sample size and greater power, we likely would have detected a significant difference between these groups of students. Because we targeted recruitment in local elementary schools, the current sample had a lower average grade level (3.25 vs. 4.23) and instructional reading level (1.90 vs. 2.80 grade equivalent) compared to the Jenkins and colleagues' sample. Further, the students in our sample were identified with a more diverse range of disabilities compared with Jenkins and colleagues' sample (see Table 1).

There were also differences in the results from both studies. First, Jenkins et al. (2017) reported overall higher decision accuracy for PM schedules than the decision accuracy of the PM schedules for the current sample's data (see Table 4). These differences in accuracy contributed to the greater statistical significance of the binomial tests Jenkins et al. conducted as well as the increased time it took each schedule to reach accuracy thresholds (see Table 5) for the current study. The increased time to reach the accuracy thresholds in the current study compared to the reported time in Jenkins et al. is perhaps the most important difference between the two studies. For example, it took most PM schedules 9 to 12 weeks to reach 70% accuracy in the current study, compared to 4 to 6 weeks for Jenkins et al. This difference has signifi-

cant practical implications regarding the number of weeks teachers need to collect CBM data before making instructional decisions.

*it took most PM schedules 9 to 12 weeks to reach 70% accuracy in the current study, compared to 4 to 6 weeks for Jenkins et al.*

It is possible that the differences in these results stem from the variability in the two study samples, particularly related to each sample's prevalence of inadequate response. A larger proportion of the current sample failed to achieve the goal rate of growth (68.63% vs. 45% of Jenkins et al.'s [2017] sample). This greater proportion of inadequate response is also reflected in the mean TG rate for this sample ( $M = 0.84$ ,  $SD = 0.55$ ) compared to the mean TG rate reported by Jenkins et al. (2017;  $M = 1.12$ ,  $SD = 0.88$ ). Despite the prevalence of inadequate response for students in the current sample, the special educators reported relatively few instructional changes for students ( $n = 21$ ; 41.18%), with four of the reported changes relating to students receiving instruction from a substitute after the teacher went on medical leave. These results are in line with previous evidence that suggests teachers struggle to adequately or effectively use CBM data to inform instruction (Stecker et al., 2005).

### Limitations

Several limitations are worth considering. First, we readministered passages once students read through the full set of available passages at their instructional level. Previous research suggests that practice effects are diminished after 10 weeks (Jenkins et al., 2005). The *t* test results indicated that TG for students reading first-grade passages, who read repeated passages prior to the 10-week mark, was not significantly different than the TG for students reading at a higher level. Even so, the risk of practice effects remains a limitation.

Second, using the same assessment data to estimate TG and weekly slopes for each PM schedule led to score overlap between the data used for the PM schedules' weekly slopes and the data used to calculate TG slopes. Score over-

lap 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 TG would be preferable. However, using a different set of passages to estimate TG introduces the additional question of equivalency of CBM passages and the comparability of student growth on those passages across vendors (see Ardoin & Christ, 2009; Ford et al., 2017).

Third, a few factors impact the generalizability of the findings. There were recruitment, attrition, and student attendance issues that impacted the final sample size. In addition, Jenkins et al. (2017) did not report the same demographic characteristics of their sample as we did (see Table 1). Therefore, we are limited in the conclusions we can draw about our findings compared to those of Jenkins and colleagues' study. Generalizability is also limited because some of the PM schedules never reached the a priori accuracy thresholds in the weeks of data collection, due to the relatively lower decision accuracy of the PM schedules for the current sample compared to the accuracy of PM schedules for students in the original study. Extending the number of weeks of data collection would have allowed for consideration of timeliness more completely.

### Next Steps

Due to the limitations, we caution against broad assertions that special educators should adopt intermittent PM schedules to ease the burden of assessment time. Research is needed to explore these research questions further. Additional research in the area of CBM will contribute to the development of specific, evidence-based criteria for CBM use in schools that address the need for both accuracy and timeliness in data-based decision making.

*Additional research in the area of CBM will contribute to the development of specific, evidence-based criteria for CBM use in schools that address the need for*

### *both accuracy and timeliness in data-based decision making.*

First, researchers should examine the potential relation between underlying prevalence of inadequate response and PM schedules' decision accuracy. Our results demonstrated lower accuracy rates across PM schedules than reported by Jenkins et al. (2017), but the lower accuracy may be due to higher rates of inadequate growth. Our sample had a much higher base rate, or prevalence, of nonresponders (i.e., students for whom intervention is not adequately working). Perhaps related to this high base rate, the majority of the errors of PM schedules in this study involved missed nonresponders. This error is problematic because the primary purpose of monitoring students' progress is to identify nonresponders and provide students with appropriately intensified instruction. It is possible that guidelines for PM schedules' accuracy criteria may need to be calibrated differently for different samples. For example, teachers in schools with high base rates of inadequate response may need to consider different PM schedules than those in schools with lower base rates of inadequate response. Depending on the student population within a school, teachers may also need to supplement CBM with other measures to increase classification accuracy of PM schedules in the same way that two-stage gated screening procedures can increase classification accuracy of screening measures (Compton et al., 2010).

These sample-specific factors related to CBM could be explored through CBM demonstration studies that manipulate base rates across large samples. Exploring the effect of prevalence would allow for a more nuanced understanding of PM schedules' adequacy in identifying student growth. These factors could also be explored through additional CBM-related replication studies. Such research could 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 ability. These data for our sample are available from the first author. Future studies could also explore the value added that supplement-

ing CBM with other measures could have in identifying nonresponsive students. Additional research in these areas would allow for a nuanced look into the factors necessary to contribute to increased accuracy of PM schedules, even for samples with high base rates of inadequate responses, such as the sample in the current study.

Second, future research should consider the ways in which PM decisions may depend on student or skill-based characteristics. An overarching goal of 1.0-WRC increase per week may not be appropriate for all students, depending on instructional reading level or severity of disability. Applying an intra-individual framework for goal setting (M. Hosp et al., 2016), which accounts for students' baseline performance levels and rates of growth in calculating individualized goal rates of growth, may be more appropriate. Future analyses could consider this alternative approach to goal setting and examine the effect it has on the accuracy of PM schedules. Future research should also explore whether results replicate across other reading CBMs (e.g., phoneme-segmentation fluency, word-reading fluency) or other academic domains (e.g., mathematics). 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.

Third, future research could focus on better understanding the amount of data needed for teachers to make decisions related to student responsiveness. Our results indicate that it takes PM schedules 9 to 12 weeks to reach accuracy thresholds, which may lead to extended exposure to ineffective instruction for students requiring instructional adaptations. Previous research has found that Tier 2 PM data may not be necessary to accurately predict first graders likely to be nonresponders to Tier 2 instruction; instead, a battery of earlier data (i.e., from screening, Tier 1, behavior ratings, and standardized measures) sufficiently predicted inadequate response (Compton et al., 2012). These results suggest the possibility of using such data to fast-track potential nonresponders immediately to Tier 3 instruction. It is

possible that a similar battery of assessment data could be used to identify nonresponders to special education interventions more rapidly and accurately than our results indicated was possible for our sample across schedules.

Last, future research should consider alternative definitions of "timeliness." We used Jenkins et al.'s (2017) definition of timeliness (i.e., the number of weeks it took different PM schedules to reach decision accuracy thresholds). Timeliness could alternatively be defined as the amount of time it takes PM schedules to identify inadequately responding students, based on TG's determination of inadequate response. This definition would serve as an index for the decision-making discrepancy between different PM schedules. This alternative timeliness index and the alternative way to assess PM schedules' effectiveness may be two ways to capture an important aspect of CBM data collection: the use of available data to make *timely* instructional changes, especially for nonresponders. Research in these areas would address the core purpose of CBM and provide educators with accurate, actionable data as soon as possible for these students.

## Conclusion

Given current initiatives to expand the use of data-based decision-making frameworks in schools (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. The findings of this study, in conjunction with the results reported by Jenkins et al. (2017), demonstrate the value of conducting replication research (i.e., documenting similarities and differences in findings across studies) and provide 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 may influence teachers' use of CBM and DBI in schools and, consequently, improve 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.
- Christ, T. J. (2006). Short-term estimates of growth using curriculum-based measurement of oral reading fluency: Estimating standard error of the slope to construct confidence intervals. *School Psychology Review, 35*(1), 128–133.
- Christ, T. J., Zopluoglu, C., Monaghan, 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*(1), 19–57. <https://doi.org/10.1016/j.jsp.2012.11.001>
- Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J. K., Barquero, L. A., Cho, E., & Crouch, R. C. (2010). Selecting at-risk first-grade readers for early intervention: Eliminating false positives and exploring the promise of a two-stage gated screening process. *Journal of Educational Psychology, 102*(2), 327–340. <http://doi.org/bsj9kr>
- Compton, D. L., Gilbert, J. K., Jenkins, J. R., Fuchs, D., Fuchs, L. S., Cho, E., Barquero, L. A., & Bouton, B. (2012). Accelerating chronically unresponsive children to tier 3 instruction: What level of data is necessary to ensure selection accuracy? *Journal of Learning Disabilities, 45*(3), 204–216. <https://doi.org/10.1177/0022219412442151>
- Conoyer, S. J., Ford, J. W., Smith, R. A., Mason, E. N., Lembke, E. S., & Hosp, J. L. (2019). Examining curriculum-based measurement screening tools in middle school science: A scaled replication study. *Journal of Psychoeducational Assessment, 37*(7), 887–898. <https://doi.org/10.1177/0734282918803493>
- 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*(4), 244–253. <https://doi.org/10.1177/0741932516648463>
- Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*(3), 219–232. <https://doi.org/10.1177/001440298505200303>
- Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education, 37*(3), 184–192. <https://doi.org/10.1177/00224669030370030801>
- 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*(4), 507–524.
- Diggs, C. R., & Christ, T. J. (2019). Investigating the diagnostic consistency and incremental validity evidence of curriculum-based measurements of oral reading rate and comprehension. *Contemporary School Psychology, 23*(2), 163–178. <http://doi.org/drgw>
- Filderman, M. J., & Toste, J. R. (2017). Decisions, decisions, decisions: Using data to make instructional decisions for struggling readers. *TEACHING Exceptional Children, 50*(3), 130–140. <https://doi.org/10.1177/0040059917740701>
- 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. <http://doi.org/gc7b44>
- Hosp, J., Ford, J., Huddle, S., & Hensley, K. (2018). The importance of replication in measurement research: Using curriculum-based measures with postsecondary students with developmental disabilities. *Assessment for Effective Intervention, 43*(2), 96–109. <https://doi.org/10.1177/1534508417727489>
- Hosp, M. K., Hosp, J. L., & Howell, K. W. (2016). *The ABCs of CBM: A practical guide to curriculum-based measurement* (2nd ed.). Guilford Press.
- Gersten, R., Compton, D., Connor, C. M., Dimino, J., Santoro, L., Linan-Thompson, S., & Tilly, W. D. (2008). *Assisting students struggling with reading: Response to intervention (RtI) and multi-tier intervention for reading in the primary grades*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. <https://ies.ed.gov/ncee/wwc/PracticeGuide/3>
- Gickling, E. E., & Armstrong, D. L. (1978). Levels of instructional difficulty as related to on-task behavior, task completion, and comprehension. *Journal of Learning Disabilities, 11*(9), 559–566. <https://doi.org/10.1177/00221947801100905>
- Individuals With Disabilities Education Act, 20 U.S.C. §§ 1400 *et seq.* (2006 & Supp. V. 2011).



- January, S.-A. A., & Ardoin, S. P. (2015). Technical adequacy and acceptability of urriculum-based measurement and the measures of academic progress. *Assessment for Effective Intervention, 41*(1), 3–15. <https://doi.org/10.1177/1534508415579095>
- 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. <https://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*(3), 144–155. <http://doi.org/gd4g8m>
- 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*(4), 213–222. <https://doi.org/10.1177/0741932516631116>
- 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*(1), 29–38.
- National Center on Intensive Intervention. (n.d.). Home page. <https://www.intensiveintervention.org>
- 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://doi.org/bk8kjk>
- Shinn, M. R., Shinn, M. M., & Langell, L. A. (n.d.). *Overview of curriculum-based measurement (CBM) and AIMSweb*. <http://www.AIMSweb.com>
- StataCorp. (2015). Stata statistical software: Release 14 [Computer software]. <https://www.stata.com>
- 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. <https://doi.org/10.1002/pits.20113>
- Treptow, M. A., Burns, M. K., & McComas, J. J. (2007). Reading at the frustration, instructional, and independent levels: The effects on students' reading comprehension and time on task. *School Psychology Review, 36*(1), 159–166.
- Yoder, P. J., Lloyd, B. P., & Symons, F. J. (2018). *Observational measurement of behavior* (2nd ed.). Brookes.

### Authors' Note

This study was registered with the Open Science Foundation (<https://bit.ly/2O8KkKX>). The research described in this article was supported in part by funding from Grant H325H140001 from the Office of Special Education Programs in the U.S. Department of Education and funding from Peabody College of Vanderbilt University. Nothing in the article necessarily reflects the positions or policies of the federal government or Vanderbilt University, and no official endorsement should be inferred.

### ORCID iD

Samantha A. Gesel  <https://orcid.org/0000-0002-8045-1607>

### Open Practices

For preregistering the plan for this research, this article earned a badge for Preregistration. The public content for the preregistration may be retrieved from <https://osf.io/udxqn/?viewonly=12a0d56a7d7c4828a62a7d46a14f3888>

### Supplemental Material

Supplemental material for this article is available online.

Manuscript received October 2019; accepted April 2020.