RMLE Online
Research in Middle Level Education

David C. Virtue, Ph.D., Editor
University of South Carolina
Columbia, South Carolina

2015 • Volume 38 • Number 6 ISSN 1940-4476

Middle Level Teachers' Perceptions of Interim Reading Assessments: An Exploratory Study of Data-based Decision Making

Deborah K. Reed, Ph.D.
Florida State University
Florida Center for Reading Research

Abstract

This study explored the data-based decision making of 12 teachers in grades 6–8 who were asked about their perceptions and use of three required interim measures of reading performance: oral reading fluency (ORF), retell, and a benchmark comprised of released state test items. Focus group participants reported they did not believe the benchmark or ORF tests accurately reflected students' comprehension abilities. Teachers held more favorable opinions of retell but admitted improvising their use of the measure rather than following mandated implementation procedures. Participants reported that only summative state assessment scores were used to plan appropriate instruction and only for large groups. Results suggest the need for improved support for data-based decision making and the development of technically adequate interim measures with relevance to the teachers expected to use them.

Keywords: interim assessment, reading, middle school, data-based decision making

The most important reasons for schools to administer reading assessments to students are to gather data for planning instruction, to evaluate the curricular program, and to determine whether students are making progress toward individual and grade level goals (Shapiro et al., 2012). These data-based decision making practices have become more common with the increase in accountability policies targeting students' reading performance (Bancroft, 2010; Hamilton et al., 2009). Although annual state assessments are one source of information that can be used to make instructional decisions, they may be ill-suited for improving teaching and learning within an academic term because they tend to be summative assessments that are administered only once—usually at the end of the year (Young & Kim, 2010).

Scriven (1967) was one of the first scholars to distinguish assessments administered for the purpose of ongoing improvement—or formative assessments—from those used to measure the end result of an educational program. In applications to student learning, formative measures allow for frequent monitoring of small steps and may help teachers and students understand and accomplish the end of year learning goal (Brookhart, 2007). However, any classroom measure with consequences for students (e.g., grade determination, placement or grouping decisions, permanent documentation of reading achievement) may make a formative assessment function more like a summative

assessment (Brookhart, 2003). Recent work defines formative assessment as a process or feedback loop that is nearly indistinguishable from the instruction itself (Hamilton et al., 2009; Wylie, 2008). Interim assessments, on the other hand, are more formally structured measures that are administered at regular intervals for the purpose of aggregating results at the classroom, grade, school, or district level (Hamilton et al., 2009).

While a wide variety of tasks may qualify as formative assessments (Bennett, 2011), the study reported here focused on teachers' perceptions and use of interim assessments that systematically measure students' reading performance. Specifically, the study sought to understand how teachers in grades six through eight perceive available data from interim assessments of students' reading performance and to document how they use that information to plan instruction.

Literature Review

Types of Interim Measures of Reading Administered in the Middle Grades

Oral reading fluency. Originally conceived as a curriculum-based measure to monitor the progress of special education students toward individual academic goals (Deno, 1985), the use of oral reading fluency (ORF) measures has expanded to encompass multiple purposes such as identifying students at risk for reading failure and predicting student performance on state accountability measures (Reschly, Busch, Betts, Deno, & Long, 2009). Teachers commonly assess ORF by having students read one or more passages aloud within a time limit while they record the errors the students make. They calculate scores based on the number of words read correctly per minute (WCPM) and they compare scores against norms to determine whether a student is meeting grade-level standards (e.g., Hasbrouk & Tindall, 2006).

Numerous studies have found a moderate to strong correlation between fluency and standardized measures of comprehension (e.g., Burke & Hagan-Burke, 2007; Spear-Swerling, 2006) as well as state reading assessments (Ditkowsky & Koonce, 2009; Espin, Wallace, Lembke, Campbell, & Long, 2010; Hunley, Davies, & Miller, 2013; Silberglitt, Burns, Madyun, & Lail, 2006). Meta-analyses confirm that ORF is a significant predictor of state assessment performance (Reschly et al., 2009; Yeo, 2010), yet teachers seem to lack confidence in WCPM scores as indicators of their students' reading comprehension (Applegate, Applegate, & Modla, 2009; Young, 2008). In measurement terms, teachers may question ORF's face validity—i.e., they may judge the instrument "on its face" as a good or bad test rather than on theoretical support for its adequacy (Fink, 1995). These concerns over ORF's face validity have prompted some test developers to add a retell component at the end of the timed reading (e.g., Good & Kaminski, 2010).

Retell. Asking a student to retell a passage is one of the most common classroom-based comprehension assessments (Cohen, Krustedt, & May, 2009). Retells regularly are included in informal reading inventories and, in the absence of more formal instruments, are a likely tool for gathering interim data in middle and high schools (Nilsson, 2008). Researchers have found that retells provide different information about students' reading skills than what is revealed by their WCPM scores alone (Kucer, 2009; Marcotte & Hintze, 2009). Moreover, adolescents' retells scored quantitatively by the number or proportion of predetermined idea units included have demonstrated moderate correlations to standardized measures of comprehension (Reed & Vaughn, 2012).

However, it is difficult to draw conclusions about the validity of retell as an interim assessment because there has been little consistency in how the measures have been administered and the responses scored across studies (Reed & Vaughn, 2012). Retells scored by the total number of meaningful words produced within a one minute time limit were found to have weaker correlations to reading comprehension scores (Bellinger & DiPerna, 2011) as have retells scored qualitatively with rubrics (Reed & Vaughn, 2012). In addition, retell scores of eighth graders lacked suitable sensitivity for differentiating students at a range of reading percentile ranks (Tindal & Parker, 1989). There is also concern that retell may not be a valid indicator of how students will perform on state test items targeting the more complex inference-making skills in the literacy standards (Reed & Vaughn, 2012). For data more closely tied to the tested standards, many schools have turned to benchmark assessments (Babo, Tienken, & Gencarelli, 2014).

Benchmark. The term *benchmark* might encompass a variety of interim assessments administered at specified intervals for comparing students' current performance to an expected level of achievement or determining whether students are on track to succeed on the summative assessment (Perie, Marion, & Gong, 2009). To distinguish *benchmark* from ORF,

the term is used here to refer to those measures intended to mimic the annual accountability test and provide information that can be used to alter classroom instruction on particular state standards for students of all ability levels (Perie, Marion, Gong, & Wurtzel, 2007). There are two different types of benchmark or standards-based assessments commonly used in the United States: those that are independently created by a commercial company (Perie et al., 2009) and those that are created locally using either original items or released versions of previously administered state tests (Quint, Sepanik, & Smith, 2008; Voloshin, 2009).

Commercial assessments are popular, in part, because they provide multiple versions for testing on a monthly or quarterly basis and also include data management software for monitoring individual students, classrooms, or schools (Marsh, Pane, & Hamilton, 2006). Yet, they have been criticized for lacking research on their efficacy and failing to involve students and teachers in meaningful improvements to learning (Babo et al., 2014; Black & Wiliam, 2007; Henderson, Petrosino, Guckenburg, & Hamilton, 2007). One notable problem is that only a few items can be included on any given test to assess each standard, making diagnosis of performance on a particular standard less reliable than overall performance (Cizek, 2007).

There are similar concerns regarding the use of released state tests. Usually states release only a limited number of versions to the public, some of which may be out of date or otherwise not aligned to current standards. In addition, states usually do not release tests with sophisticated data management tools for monitoring student progress over time. The quality of these tools is an important factor in determining the extent to which teachers use interim assessments for instructional purposes (Young & Kim, 2010).

Despite these concerns, different varieties of benchmarks seem to have strong face validity among educational administrators. In a study by Marsh and associates (2006), 80% of superintendents and 80% of principals reported that benchmark assessments were more useful than state tests and that they were moderately to very useful in guiding instructional decisions. However, because teachers in the study preferred classroom-based tests, there may be deeper issues involving teachers' perceptions of the assessments.

Teacher Perceptions of Interim Assessments

While interim data on students' reading performance is widely available, teachers do not consistently use this information in meaningful ways. Teachers may be somewhat resistant to ongoing interim assessments (Black & William, 1998), and the perceived face validity of instruments is often in contrast to their psychometric properties. For example, ORF has demonstrated strong correlations to state accountability tests but reportedly has low face validity among teachers and reading specialists as an indicator of students' comprehension abilities (Applegate et al., 2009). Conversely, there is a paucity of research supporting the use of benchmarks to improve student outcomes, yet they are widely used because they are believed to provide information relevant to teachers (Henderson et al., 2007).

A meta-analysis of systematic interim evaluation found that providing teachers with explicit guidance on and distinct processes for the use of data in instructional decision making resulted in higher effects on student achievement than allowing teachers to make their own decisions about how to interpret and use the data (Fuchs & Fuchs, 1986). Similarly, a narrative review suggested that teachers' perceptions and use of interim data may hinge on the quality and extent of the training they receive as well as on the depth of their instructional knowledge (Young & Kim, 2010).

Unfortunately, not all efforts to improve teachers' data-based decision making are successful. Findings from a yearlong study of elementary school teachers engaged in a district-wide interim assessment initiative indicate the teachers did not use the data to substantially change their teaching or testing practices (Goertz, Nabors Olah, & Riggan, 2009). The general education teachers in the study were expected to use assessment information to differentiate instruction—to plan enrichment for students who were excelling and interventions for students who were struggling. Some researchers have noted teachers use assessment results merely to validate their previously formed impressions of students' abilities (Nabors Olah, Lawrence, & Riggan, 2010) or to dichotomously group middle level students as readers who struggle and readers who do not struggle, treating those within each broad group as homogeneous (Moreau, 2014). Others have found that even when teachers agree that improving their ability to use data would aid their professional growth, they continue to believe their own classroom-based tests are better indicators of students' learning (Wayman, Cho, & Johnston, 2007).

Purpose of the Study

Much of the extant literature about formative assessment in reading and data-based decision making has focused on elementary school settings. Less is known about how these assessments might inform teaching and learning among adolescents who have increasingly high stakes attached to their reading performance (Green et al., 2008). Therefore, this exploratory study sought to better define middle level teachers' perceptions of interim assessments administered for the purposes of data-based decision making. The research focused on the middle grades for two reasons. First, the numbers of students identified with reading disabilities nearly doubles in early adolescence (Leach, Scarborough, & Rescorla, 2003; Lipka, Lesaux, & Siegel, 2006), suggesting an important role for data-based decision making in these grades. Second, middle level schools must address the unique developmental needs of young adolescent students (National Middle School Association [NMSA], 2010), and they have a critical role in ensuring students are prepared for advanced academic literacy associated with higher educational attainment levels (ACT, 2008). The research question that guided the study was: How do teachers in grades six through eight perceive available interim assessment data on students' reading performance and use that information to plan instruction?

Method

Participants

The two participating middle level schools were located in different cities of the southwestern United States, more than 400 miles apart. A majority of the students at both campuses were Hispanic (M = 61%) and received free or reduced price lunch (M = 53%). Schools with similar demographics have placed added emphasis on monitoring all students' progress toward grade level standards (Bancroft, 2010).

The principal investigator (PI) conducted focus group interviews with 12 teachers from two sites (Site 1 = 3 teachers; Site 2 = 9 teachers). Because the research concerned the use of data-based decision making for students regardless of their ability levels, all general education English language arts/reading (ELAR) teachers were invited to share their perceptions of interim assessments and consented to do so. The focus group interviews at Site 2 also included the campus literacy coach, one teacher who taught English as a Second Language (ESL) in addition to ELAR, and one teacher who taught both social studies/humanities and ELAR. The teachers' years of experience ranged from 2 to 20 years. Demographic data are provided in Table 1.

Table 1
Teacher Demographics

Characteristic	Site 1 N = 3	Site 2 N = 9
Female	3	6
Male	0	3
White	3	1
Hispanic	0	8
ELAR grade 6	1	2
ELAR grade 7	1	2
ELAR grade 8	1	2
ESL/ELAR grades 6–8	0	1
Social Studies/Humanities + ELAR (grades 7–8)	0	1
Literacy Coach	0	1

Note. ELAR = English language arts and reading; ESL = English as a second language

Measures

Focus group participants were asked questions about three interim measures—ORF, retell, and benchmark—and the summative state assessment of reading. District or state policy mandated that schools administer all tests. The summative measure was included to provide a context for understanding teachers' remarks about the interim measures. That is, because the latter are intended to predict performance on the former, teachers' understanding and beliefs about the state assessment could influence their understanding and beliefs about the ORF, retell, and benchmark tests.

Oral reading fluency (ORF). ORF was assessed with the Texas Middle School Fluency Assessment ([TMSFA]; Texas Education Agency, University of Houston, & The University of Texas System, 2010). This measure was developed with a large and diverse sample of students in grades six through eight, representative of those below, at, and above grade level performance. The test involves students reading a series of three passages aloud for one minute each to determine the number of words read correctly per minute. Passages are designated by grade and testing point (i.e., beginning of year, middle of year, end of

year), comprised of narrative and expository texts, and presented in successive levels of difficulty. Test-retest reliabilities ranged from 0.88 to 0.93, and intra-class correlation coefficients ranged from 0.88 to 0.91. Developers reported the instrument demonstrated moderate to strong correlations with standardized measures of reading ($r = 0.57 – 0.67$). The criterion validity was established with the state reading test ($r = 0.50$).

Retell. After students completed each one minute reading in the TMSFA, they were prompted to produce a retell with, "Can you tell me everything you remember reading in the passage?" Each time they paused, students were prompted to continue retelling with, "Do you remember anything else?" until they indicated they could recall no more information. The prompts were intended to elicit as much information from the student as possible because retells were scored by the proportion of pre-determined idea units retold out of the total number read. Testers transcribed each student's response as it was being delivered, and the transcription was later compared against a list of idea units developed for each passage. The number of identified idea units each student actually retold was divided by the maximum idea units possible for his/her total word count, resulting in a percentage score. The scoring instrument demonstrated a strong intra-class correlation (i.e., 0.98) that suggested inter-rater agreement would not have occurred by chance (Reed, Vaughn, & Petscher, 2012). The retell measure was validated through confirmatory factor analysis by comparing the fit [χ^2 (32) = 97.316; CFI = 0.96; TLI = 0.94; RMSEA = 0.08] of a three factor model of reading to the data from a diverse sample of seventh and eighth graders (Reed, Vaughn, & Petscher, 2012). It had weak but significant ($p < .01$) correlations to the state reading test ($r = 0.26$) and standardized measures of reading comprehension ($r = 0.16$-0.21).

Released-test benchmark. Personnel from the two districts created the benchmark test for each school by using released state assessment items. No information was available on the technical adequacy of using an assemblage of released items as an interim assessment, but the locally created measures demonstrated a strong correlation ($r \cong 0.68$) to the state test. This robust relationship might be expected with what was essentially a newer version of the benchmark. Similarly, a hierarchical regression revealed the district benchmarks accounted for 40% to 53% of the unique variance on the state test.

Procedures

The research team interviewed teachers in focus groups at both sites. This format was selected over one-on-one interviews because it was believed that peer interaction would be valuable in challenging the thinking of participants; helping to identify potentially conflicting opinions; and stimulating richer, co-constructed insights than would have resulted from interviews (Kitzinger, 1995). For instance, some questions related to the way teachers' made sense of the testing protocols and, therefore, provided insight into their improvisations. Hence, it was important to see how teachers described this collectively. The research team also believed that other questions would become redundant if they asked them of teachers individually. The focus group format gave teachers time to reflect and the ability to use colleagues' statements as a stimulus for further responses, including more critical comments than might have been offered in individual interviews (Watts & Ebbutt, 1987).

The three teachers at Site 1 preferred to meet with the PI and a research assistant (RA) during a common planning period, and the nine teachers (including the literacy coach) at Site 2 preferred to meet after school. Both sessions were held in a classroom with all participants and the PI sitting at desks arranged in a circle. The PI served as moderator and, per the recommendations of Kidd and Parshall (2000), the RA sat off to the side to record the order of speakers and any significant nonverbal behavior. In addition, the focus groups were audio recorded to verify the accuracy of field notes, capture comments from multiple speakers at a time, and weave together verbal and nonverbal data the PI and RA collected.

After explaining the purpose of the focus group and ensuring the confidentiality of the information shared, the PI asked the questions in the approved protocol:

1. When you have access to state assessment, ORF, retell, and benchmark data for a student, which do you think provides the most accurate information about the student's reading comprehension abilities?

2. How do you use state assessment data to plan your instruction?

3. How do you use ORF data to plan your instruction?

4. Do you think data from the retell measure is an important part of understanding your students' reading abilities and planning your instruction?*

5. If you were required to administer one reading assessment of your choice three times per year, which assessment would you choose: the state assessment, ORF, retell, or benchmark?

6. Is there anything we did not cover today that you would like to add so I better understand how you perceive and use reading assessment data?

The questions were written specifically to elicit information about how teachers perceived and used data from the assessments included in the study and, thus, were considered a *priori* categories of interest. To avoid influencing responses, the PI did not offer reaction to any statements and encouraged participants to talk to one another. Before moving to the next question in the protocol, the PI verbally reviewed the notes recorded for the current question to obtain participants' confirmation of its accuracy. Given the difficulties of reconvening focus groups, the research team decided to conduct member checking in real time while the group was being conducted, as Kidd and Parshall (2000) recommended.

The smaller focus group interview lasted about 40 minutes and the larger one lasted about 60 minutes. In the post-session debriefs, the PI and RA discussed their impressions, observations of nonverbal behavior, and any concerns from potential imposition of group norms (Kitzinger, 1995).

Analysis

The PI and an independent consultant analyzed the field notes and transcribed audio recordings. In the first phase of analysis, the focus group questions were used as broad categories within which points of view at the individual and group levels could be identified (Kidd & Parshall, 2000). The points of view were further discriminated into areas of agreement and disagreement with a particular focus on identifying disconfirming evidence or suggestions (e.g., data on nonverbal behavior) that alternative viewpoints might have been suppressed (Carey & Smith, 1994). Using Morgan's (1997) suggestions, the issues were examined to determine whether participants expressed them in one or both focus groups, returned to them multiple times, or raised them spontaneously. The researcher and consultant gathered supporting quotations for the areas of agreement and disagreement and discussed them to resolve any questions about how they categorized interview content and to return to the data to continue the analysis. In one instance, the researcher contacted a teacher to seek clarification of a point she made during the interview.

Last, the researcher and consultant reviewed responses for repetition across the broad categories/questions. When they agreed that responses to two questions had extensive overlap, the categories were combined. The final categories were renamed with more descriptive titles:

- Perceptions of most commonly used data (responses from questions 1 and 2)
- Perceptions of less commonly used data (responses from questions 3 and 4)
- Preferences for required testing (responses from question 5)
- Training and improvisation (responses from question 6)

Results

Perceptions of Most Commonly Used Data

At both sites, participants indicated that the summative state reading test results were used more often than any interim assessment data when planning instruction. They described the particular use of the state assessment results in the aggregate: identifying the "weakest" objectives (i.e., those on which one or more grade levels had the lowest percentage of correct responses) in order to prioritize instruction in those skills. A teacher at Site 1 explained, "For example, we really hammered *tone* and *mood* this year." An ELAR teacher at the other site stated, "*Summarization* gets us, so we do activities decided on by the department to target that objective."

Teachers at neither site described using disaggregated results or individual student performance for planning small group or differentiated instruction. In addition, no teachers mentioned using data from the benchmark to plan instruction because the teachers did not believe the results accurately reflected student performance. A participant at Site 1 offered this explanation:

> Students bring their "A-game" on [state testing] day, so you can really understand what they are capable of. When it's the benchmark, students have the attitude,"It's just the benchmark, so who cares?" Those results aren't as reflective of their capabilities.

* The wording of this question differed from questions 2 and 3 because retell was included at the end of the ORF assessment in the TMSFA but was not required by the state for diagnostic purposes.

Site 2 administered the benchmark six times per year, so teachers reacted to the frequency of testing. One teacher said, "I think the students are 'benchmarked' to death. That makes the results not accurate at all." Different teachers commented, "We don't want the benchmarks," at different times during the focus group. Nevertheless, one participant acknowledged, "I like that it gives a snapshot of where students are at, but not so many!" When this teacher suggested a reduction to administering the district benchmark two times per year, three of her colleagues nodded in support. Another added, "We get students from all over because of the [military] base, so not all families come with [our state test] scores. We need to know how those students are doing."

Site 2 participants suggested they were more apt to rely on teacher generated data. This was initially brought up by the ESL/ELAR teacher, who seemed to dismiss the interim assessments specifically: "We use the exams you mentioned because they are required, but my own questions guide me. They tell me what each student needs." The department chair echoed this idea, emphasizing that their ELAR curriculum was organized around reading novels that made teachers "attune to the red flags. We know right away that [the students] are not understanding." Another ELAR teacher stated, "In class, with the teacher directed back-and-forth, we can figure out what's going on. We don't need the tests." Because the use of teacher questioning as a form of assessment was not included in this study of structured interim assessments, there were no focus group items probing for information on this type of data. No one at Site 1 spontaneously volunteered similar information.

Perceptions of Less Commonly Used Data

The schools employed different models for gathering interim ORF data, which is consistent with the findings of other studies (Munir-McHill, Bousselot, Cummings, & Smith, 2012). Only the seventh grade ELAR teacher at Site 1 and the literacy coach at Site 2 were administering the ORF and retell measures. The teachers knew the tests could be used with all students, but they were administering the measures in compliance with a state legislative mandate regarding students who failed the state reading test. Participants reported that the results were used to confirm the placement of students in a reading intervention class in which the curriculum was delivered by a computer program with its own diagnostic assessment. Despite its limited implementation at the time, all 12 teachers at both research sites unanimously agreed that the retell measure would be an important part of understanding all their students' reading abilities—not just those who failed the state reading test. As a teacher at Site 1 put it, "Yes, definitely because a kid can fly through the words but, then, can't understand it." Her colleagues emphatically agreed. The face validity of retell seemed predicated on two beliefs expressed in the following statements: "It's a lot like what we do in the classroom;" and "It's close to [the state reading test]."

Preferences for Required Testing

The teachers' support of the retell assessment was apparently strong. Even though only 2 of the 12 focus group participants were actually administering the ORF and retell assessments as part of the legislative mandate, nine teachers across the sites indicated they would prefer to give those combined tests three times per year rather than any other assessment. At Site 2, the literacy coach stated her choice simply: "The [combined ORF and retell measure] because we're familiar with it." Three of her colleagues, who were not as familiar with the measures, instead preferred the less frequently administered benchmark or no test at all, just teacher directed questioning.

At Site 1, all three teachers were quick to respond with their preference, "Not [the state assessment]!" The eighth grade teacher, who had not been administering the ORF or retell measures, offered elaboration: "I would like the [combined ORF and retell measure] because it's a quick way to assess students and allows us to look at growth over the year." The teachers at that site seemed very interested in having an assessment with "different versions" or forms that could be used to chart student progress. Nevertheless, they were concerned about the amount of time it took to individually administer the tests. One teacher asserted:

> In an ideal world, [the state department of education] would provide us the time and resources so we could do the [ORF and retell tests], but not for every student. I don't think it would make any difference for the strong kids, but just for the kids who struggle.

Training and Improvisation

The three teachers at Site 1 emphasized the need for more thorough training. The seventh grade teacher who was administering the combined ORF and retell measure for her school stated, "You can't just come in and look at it and figure it out. The training was necessary." Her colleague added, "Teachers want to know that they are doing it right." In contrast, the teachers at Site 2 made several statements that

indicated they had low fidelity to the mandated test administration procedures. For example, rather than transcribing the entire retell as the directions indicated, the literacy coach (who was also a trainer for the ORF and retell measures) said, "I started just writing 'dot, dot, dot' when they went on and on, and I knew 'this little guy got it; he understands.' I didn't write down everything he said." Although her colleagues were not officially responsible for administering the mandated ORF and retell tests, they were trained to do so and stated that they applied the same procedures in their classroom fluency practice. However, two teachers made statements that also indicated improvisation. One stated, "I have the students give just three things: one from the beginning, the middle, and the end. They gave so much otherwise!" Another shared, "Yeah, I ask them to tell in one sentence, to summarize for me." Even after the literacy coach reminded her colleagues that they had to "ask it the way it is on the test," one teacher said, "Well, one time I did tell them, 'Tell me in 10 words.'"

Discussion

This study investigated middle level teachers' perceptions of and use of interim measures of reading performance. Consistent with the findings of McMillan (2003), the study revealed tension between teachers' knowledge, beliefs, and expectations of assessments and testing policies mandated at the district or state level that seemed to discourage data-based decision making. This tension is discussed in the sections that follow with respect to each type of interim measure considered in the study.

Oral Reading Fluency

A single, designated teacher at each site administered the ORF measure for the purposes of complying with a legislative mandate and confirming that the students who failed the state reading assessment should be placed in an intervention. A computer program then designed, delivered, and monitored the instruction in the intervention class, thus alleviating any compulsion for teachers to further explore the ORF data. Teachers also did not use the ORF measure to further screen students who passed the state assessment but failed a benchmark test. These practices were similar to the findings of previous research in which middle level teachers used assessment results to broadly determine which students were struggling with reading and which were not (Moreau, 2014). Therefore, there was potential for overreliance on teacher judgment rather than objective measures for ongoing identification of students who could benefit from supplemental reading intervention (Madelaine & Wheldall, 2005), particularly at Site 2 where focus group participants described reluctance to use the interim assessments. Consistent with other research at the middle level (Hunley et al., 2013), the ORF measure used at these two sites had a moderate correlation to the state assessment. As reported elsewhere (Young, 2008), the focus group participants in the present study doubted that students' ORF was truly indicative of their comprehension, but the teachers still expressed interest in having a tool for monitoring students' progress over the year—as long as the retell component was included.

Retell

Teachers believed the retell instrument, which had the weakest concurrent validity with the state reading assessment, provided valuable information about students' comprehension because it was related to classroom instruction and to the state test items. It may be that teachers' lack of fidelity to or familiarity with the specified retell administration and scoring protocols mistakenly led them to draw more positive conclusions about retell than about the other interim assessments. For example, the comments of teachers at Site 2 indicated their instructional applications of retell were not aligned to the prompting procedures outlined in the assessment manual. As found in previous research (Reed & Petscher, 2012; van den Broek, Tzeng, Risden, Trabasso, & Basche, 2001), the focus groups remarked that altering the prompt changed the quality and quantity of student responses. In fact, that was the rationale for making such significant changes. The teachers all said that their alterations were instructionally sound and reflective of a student's reading comprehension, even though they created a lack of consistency in how retell was defined and measured from classroom to classroom.

Benchmark

Teachers seemed to only value the district benchmark test—the interim assessment with the most robust relationship to the state reading assessment—for giving a "snapshot" of student performance. They did not report using it strategically to plan instruction for objectives with which students seemed to have more difficulty. This was true despite the fact that teachers reportedly examined the official state assessment results by objective to set annual grade-level or school-wide instructional priorities at both sites. Goertz and associates (2009) suggested a benchmark could be

used to inform what to teach but not how to teach it. However, the middle level teachers interviewed for this study did not acknowledge changing the content (i.e., what) or the methods (i.e., how) of their instruction based on interim assessment results that were aligned with the state accountability measure.

The teachers' lack of examination of the benchmark data may be due to the fact that they had fewer resources available to them to use these results compared to the state assessment data. At the time of this study, the state department of education provided annual assessment reports with student passing rates by objective and standard. In addition, there were item analyses showing response rates on individual items tied to individual standards. To obtain the same information from the benchmarks, teachers would have to calculate the passing rates by objective and determine which items were aligned to standards that had not yet been taught. This kind of unstructured and unsupported use of interim assessments may prevent educators from using them as a catalyst for instructional improvement (Young & Kim, 2010).

Another reason focus group teachers may not have used benchmark scores was because they lacked confidence in them. They believed that students did not take the test seriously or that the test was administered too frequently to be accurate. At Site 2, which "benchmarked" every six weeks, teachers' frustration with the test was great enough for them to express a desire not to use any interim assessments at all. Given that previous studies have called into question the reliability and predictive validity of benchmarks (Babo et al., 2014; Cizek, 2007), it is probable the teachers' perceptions were not unfounded. Hence, it is understandable why teachers at that site found their classroom-based judgments of student performance were more helpful and informative, a finding that is consistent with previous research (Marsh et al., 2006; Wayman et al., 2007).

Data-based Decision Making

The intent of this exploratory study was to determine teachers' perceptions and uses of interim measures in a data-based decision making environment, but the focus group only yielded information on the former. That is, participants generally did not use interim data in a systematic manner to inform teaching and learning. There are two possible explanations for this. First, while data-based decision making is recommended for planning appropriate instruction for all students (Hamilton et al., 2009), teachers in the focus groups may have found the investment in data analysis less worthwhile when working with students who represent all ability levels rather than just the lower end of the spectrum. One teacher at Site 1 hinted at this when she expressed support for using the combined ORF and retell measure but only with the students who struggled. For the others, she thought it would not "make any difference." If that is true, it points to the need for measures that better align to the relevant skills that will help students at higher and lower levels of ability continue to make progress. By some definitions, formative assessment would be more appropriate in this regard because it is more seamlessly integrated with the cycle of teaching and learning than interim assessment (Brookhart, 2003; Hamilton et al., 2009).

Second, it is also possible that the teachers' remarks were a demonstration of confirmation bias (Nickerson, 1998). In other words, they may have been willing to consider the data that supported their own notions about teaching and learning, but they would not look for counter evidence (Mercier & Sperber, 2011). In fact, the teachers sought to rebut negative arguments. For example, they used the retell instrument to confirm that the students already identified as needing a reading intervention really belonged there, but the teachers did not actively screen for other students whose difficulties might have arisen during the year. They also justified improvising retell and ignoring the benchmark test because they believed that their own abilities to gauge student progress were more accurate than interim test data. It is likely that a greater level of support would be necessary to overcome confirmation bias—if it actually existed—and to help teachers critically examine their own beliefs (McHatton, Parker, & Valice, 2013).

Limitations and Directions for Future Research

Although the research team deemed focus group interviews of teachers to be an appropriate strategy for this study, the group setting could potentially have increased an individual participant's tendency to provide input consistent with group norms (Carey & Smith, 1994). Every attempt was made to elicit the comments of each participant and analyze the data for suggestions that alternative points of view were suppressed. There were several remarks that were not affirmations of colleagues' statements, but it is impossible to know if more divergent responses would have been expressed in individual interviews. In addition, the sample was relatively small and more

representative of general education teachers than of interventionists for students with identified reading difficulties. Because reading intervention at the sites included in this study was planned and delivered by a computer, there was a reduced role for teachers in making instructional decisions for the students with the lowest performance.

Due to the data collection model employed by the campuses (Munir-McHill et al., 2012), only two of the 12 teachers interviewed were directly involved in the administration and scoring of the ORF and retell measures as part of the school's regular testing schedule. Comments might have been different if elicited from teachers who taught a higher percentage of adolescents struggling with reading or who were more directly involved in administering all the interim assessments included in this study. Often, participants' responses relied on assumptions about the ORF and retell assessments rather than in-depth experiences with them. Discussion during the focus group seemed to stimulate some participants' interest in the instrument, so it is possible that the training and support they requested could lead to different beliefs or confidence in the results. Previous research has also documented pre- and in-service teachers' desire and need for more thorough assessment training (Young & Kim, 2010) as well as professional development about how to help middle level students who struggle with reading (Moreau, 2014).

Practical Implications

The middle grades are critical years for student development (ACT 2008; NMSA, 2010). If the goal of administering interim assessments to young adolescents is to guide instructional decisions and appropriately challenge students of varying ability levels (Perie et al., 2009), the instruments must provide accurate data that will be acceptable and meaningful to middle level teachers (Shute, 2007). The results of this study demonstrate that none of the three interim assessments investigated met these parameters. The instrument with the strongest technical adequacy, the district benchmark, was not used by the focus group teachers because they did not believe it was an accurate reflection of students' reading comprehension abilities. In contrast, retell had the lowest concurrent validity with the state assessment but the most favorable opinion among teachers. Nevertheless, the understanding and use of retell data was inconsistent and based more on how teachers implemented retelling a passage in their classroom instruction than on mandated procedures for administering the assessment. Whether or not more process-oriented formative assessment can resolve these issues in the middle grades should be explored in future research (Bennett, 2011).

With respect to assessment policy, the tests considered in this study were all mandated by the district or state. Even though the intent of the interim assessment policies was to provide actionable information, mandating the tests did not guarantee data-based decision making would take place (Marsh et al., 2006; Perie et al., 2007). Recall that participants reported using the summative state reading test results for nothing more than large group (i.e., grade level or school-wide) planning of *what* to teach. Therefore, improving students' reading abilities across performance categories may hinge on providing greater support to teachers and on implementing measures that are well aligned to the relevant skills important for the population in question. This, in conjunction with long term professional development in assessment and instruction (Shepard, 2000), may help teachers understand how they can effectively use the tests as a resource and a complement to their professional judgments.

References

ACT. (2008). The forgotten middle: Ensuring that all students are on target for college and career readiness before high school. Iowa City, IA: Author. Retrieved from www.act.org

Applegate, M. D., Applegate, A. J., & Modla, V. B. (2009). "She's my best reader; she just can't comprehend": Studying the relationship between fluency and comprehension. *The Reading Teacher, 62*, 512–521. doi: 10.1598/RT.62.6.5

Babo, G., Tienken, C. H., & Gencarelli, M. A. (2014). Interim testing, socio-economic status, and the odds of passing grade 8 state tests in New Jersey. *Research in Middle Level Education Online, 38*(3), 1–9. Retrieved from www.amle.org/portals/0/pdf/rmle/rmle_vol38_no3.pdf

Bancroft, K. (2010). Implementing the mandate: The limitations of benchmark tests. *Educational Assessment, Evaluation and Accountability, 22*, 53–72. doi: 10.1007/s11092-010-9091-1

Bellinger, J. M., & DiPerna, J. C. (2011). Is fluency-based story retell a good indicator of reading comprehension? *Psychology in the Schools, 48*, 416–426. doi: 10.1002/pits.20563

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy and Practice, 18*, 5–25. doi: 10.1080/0969594X.2010.513678

Black, P., & Wiliam, D. (1998). *Assessment and classroom learning. Assessment in Education: Principles, Policy and Practice, 5*, 7–74. doi: 10.1080/0969595980050102

Black, P., & Wiliam, D. (2007). A theoretical foundation for formative assessment. In J. H. McMillan (Ed.), *Formative classroom assessment: Theory into practice*. (pp. 29–42). New York, NY: Teachers College Press.

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice, 22*, 5–12. doi: 10.1111/j.1745-3992.2003.tb00139.x

Brookhart, S. M. (2007). Expanding views about formative classroom assessment: A review of the literature. In J. H. McMillan (Ed.), *Formative classroom assessment: Research, theory and practice* (pp. 43–62). New York, NY: Teachers College Press.

Burke, M. D., & Hagan-Burke, S. (2007). Concurrent criterion-related validity of early literacy indicators for middle of first grade. *Assessment for Effective Intervention, 32*, 66–77. doi: 10.1177/1535084070320020401

Carey, M. A., & Smith, M. W. (1994). Capturing the group effect in focus groups: A special concern in analysis. *Qualitative Health Research, 4*, 123–127. doi: 10.1177/104973239400400108

Cizek, G. J. (2007). Formative classroom assessment and large-scale assessment: Implications for future research and development. In J. A. McMillan (ed.), *Formative classroom assessment: Theory into practice* (pp. 99–115). New York, NY: Teachers College Press.

Cohen, L., Krustedt, R. L., & May, M. (2009). Fluency, text structure, and retelling: A complex relationship. *Reading Horizons, 49*, 101–124.

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

Ditkowsky, B., & Koonce, D. A. (2009). Predicting performance on high-stakes assessment for proficient students and students at risk with oral reading fluency growth. *Assessment for Effective Intervention, 35*, 159–167. doi: 10.1177/1534508409333345

Espin, C., Wallace, T., Lembke, E., Campbell, H., & Long, J. D. (2010). Creating a progress-monitoring system in reading for middle-school students: Tracking progress toward meeting high-stakes standards. *Learning Disabilities Research & Practice, 25*, 60–75. doi: 10.1111/j.1540-5826.2010.00304.x

Fink, A. (1995). *How to measure survey reliability and validity* (Vol. 7). Thousand Oaks, CA: Sage.

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children, 53*, 199–208.

Goertz, M. E., Nabors Olah, L., & Riggan, M. (2009, December). Can interim assessments be used for instructional change? *CPRE Policy Briefs: Reporting on Issues and Research in Education Policy and Finance*, RB-51.Philadelphia, PA: Consortium for Policy Research in Education. Retrieved from: www.cpre.org

Good, R. H., & Kaminski, R. A. (2010). *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.). Eugene, OR: Dynamic Measurement Group, Inc. Retrieved from: http://www.dibels.org.

Green, W. L., Caskey, M. M., Musser, P. M., Samek, L. L., Casbon, J., & Olson, M. (2008). Caught in the middle again: Accountability and the changing practice of middle school teachers. *Middle Grades Research Journal, 3*(4), 41–72.

Hamilton, L., Halverson, R., Jackson, S., Mandinach, E., Supovitz, J., & Wayman, J. (2009). *Using student achievement data to support instructional decision making* (NCEE 2009-4067). 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/.

Hasbrouk, J., & Tindall, G. A. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *The Reading Teacher. 59,* 636–644. doi: 10.1598/RT.59.7.3

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2007). *Measuring how benchmark assessments affect student achievement* (Issues & Answers Report, REL 2007–No. 039). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast and Islands. Retrieved from http://ies.ed.gov/ncee/edlabs.

Hunley, S. A., Davies, S. C., & Miller, C. R. (2013). The relationship between curriculum-based measures in oral reading fluency and high-stakes tests for seventh grade students. *Research in Middle Level Education Online, 36*(5), 1–8. Retrieved from www.amle.org/portals/0/pdf/rmle/rmle_vol36_no5.pdf

Kidd, P. S., & Parshall, M. B. (2000). Getting the focus and the group: Enhancing analytical rigor in focus group research. *Qualitative Health Research, 10*, 293–308. doi: 10.1177/104973200129118453

Kitzinger, J. (1995, July 29). Qualitative research: Introducing focus groups. *British Medical Journal, 311*, 299–302. PMCID: PMC2550365

Kucer, S. B. (2009). Examining the relationship between text processing and text comprehension in fourth-grade readers. *Reading Psychology, 30*, 340–358. doi: 10.1080/02702710802411604

Leach, J. M., Scarborough, H. S., & Rescorla, L. (2003). Late-emerging reading disabilities. *Journal of Educational Psychology, 95*, 211–224. doi: 10.1037/0022-0663.95.2.211

Lipka, O., Lesaux, N. K., & Siegel, L. S. (2006). Retrospective analysis of the reading development of grade 4 students with reading disabilities: Risk status and profiles over 5 years. *Journal of Learning Disabilities, 39*, 364–378. doi: 10.1177/00222194060390040901

Madelaine, A., & Wheldall, K. (2005). Identifying low-progress readers: Comparing teacher judgment with a curriculum-based measurement procedure. *International Journal of Disability, Development, and Education, 52*, 33–42. doi: 10.1080/10349120500071886

Marcotte, A. M., & Hintze, J. M. (2009). Incremental and predictive utility of formative assessment methods of reading comprehension. *Journal of School Psychology, 47*, 315–335. doi: 10.1016/j.jsp.2009.04.003

Marsh, J. A., Pane, J. F., & Hamilton, L. S. (2006). *Making sense of data-driven decision making in education.* Santa Monica, CA: Rand Corporation.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences, 34*, 57–111. doi: 10.1017/S014025X10000968

McHatton, P. A, Parker, A. K., & Valice, R. K. (2013). Critically reflective practitioners: Exploring our intentions as teacher educators. *Reflective Practice, 14*, 392–405. doi: 10.1080/14623943.2013.767235

McMillan, J. H. (2003). Understanding and improving teachers' classroom assessment decision making: Implications for theory and practice. *Educational Measurement: Issues and Practices, 22*, 34–43. doi: 10.1111/j.1745-3992.2003.tb00142.x

Moreau, L. K. (2014). Who's really struggling? Middle school teachers' perceptions of struggling readers. *Research in Middle Level Education Online, 37*(10), 1–17. Retrieved from www.amle.org/portals/0/pdf/rmle/rmle_vol37_no10.pdf

Morgan, D. L. (1997). *Focus groups as qualitative research* (2nd ed.). Thousand Oaks, CA: Sage.

Munir-McHill, S., Bousselot, T., Cummings, K. D., & Smith, J. L. M. (2012). *Profiles in school-level data-based decision making.* Paper presented at the annual meeting of the National Association of School Psychologists, Philadelphia, PA.

Nabors Olah, L., Lawrence, N. R., & Riggan, M. (2010). Learning to learn from benchmark assessment data: How teachers analyze results. *Peabody Journal of Education, 85*, 226–245. doi: 10.1080/01619561003688688

National Middle School Association. (2010). *This we believe: Keys to educating young adolescents.* Westerville, OH: Author.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomena in many guises. *Review of General Psychology, 2*, 175–220.

Nilsson, N. L. (2008). A critical analysis of eight informal reading inventories. *The Reading Teacher, 61*, 526–536. doi: 10.1598/RT.61.7.2

Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practices, 28*(3), 5–13.

Perie, M., Marion, S., Gong, B., & Wurtzel, J. (2007). *The role of interim assessments in a comprehensive assessment system: A policy brief.* Washington, DC: The Aspen Institute.

Quint, J. C., Sepanik, S., & Smith, J. K. (2008). *Using student data to improve teaching and learning: Findings from an evaluation of the formative assessments of students thinking in reading (FAST-R) Program in Boston elementary schools.* New York, NY: MDRC. Retrieved from http://www.mdrc.org/publications/508/full.pdf

Reed, D. K., & Petscher, Y. (2012). The influence of testing prompt and condition on middle school students' retell performance. *Reading Psychology, 33*, 562–585. doi: 10.1080/02702711.2011.557333

Reed, D. K., & Vaughn, S. (2012). Retell as an indicator of reading comprehension. *Scientific Studies of Reading, 16*, 187–271. doi: 10.1080/10888438.2010.538780

Reed, D. K., Vaughn, S., & Petscher, Y. (2012). The validity of a holistically-scored retell protocol for determining the reading comprehension of middle school students. *Learning Disability Quarterly, 35*, 76–89. doi: 10.1177/0731948711432509

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 correlational evidence. *Journal of School Psychology, 47*, 427–469. doi: 10.1016/i.jsp.2009.07.001

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39–83). Chicago, IL: Rand McNally.

Shapiro, E., Hilt-Panahon, A., Gischlar, K., Semeniak, K., Leichman, E., & Bowles, S. (2012). An analysis of consistency between team decisions and reading assessment data within an RTI model. *Remedial & Special Education, 33*, 335–347. doi: 10.1177/0741932510397763

Shepard, L. (2000). *The role of classroom assessment in teaching and learning.* Technical Report No. 517. Boulder, CO: National Center for Research on Evaluation, Standards, and Student Testing.

Shute, V. J. (2007). *Focus on formative feedback.* ETS Research Report. Retrieved from http://www.ets.org/research/contact.html.

Silberglitt, B., Burns, M. K., Madyun, N. I. H., & Lail, K. E. (2006). Relationship of reading fluency assessment data with state accountability test scores: A longitudinal comparison of grade levels. *Psychology in the Schools, 43*, 527–535. doi: 10.1002/pits.20175

Spear-Swerling, L. (2006). Children's reading comprehension and oral reading fluency in easy text. *Reading & Writing: An Interdisciplinary Journal, 19*, 199–220. doi: 10.1007/s11145-005-4114-x

Texas Education Agency, University of Houston, & The University of Texas System. (2010). *Texas middle school fluency assessment.* Austin, TX: Author.

Tindal, G., & Parker, R. (1989). Development of written retell as a curriculum-based measure in secondary programs. *School Psychology Review, 18*, 328–343.

van den Broek, P., Tzeng, Y., Risden, K., Trabasso, T., & Basche, P. (2001). Inferential questioning: Effects on comprehension of narrative texts as a function of grade and timing. *Journal of Educational Psychology, 93*, 521–592. doi: 10.1037//0022-0663.93.3.521

Voloshin, D. (2009). *An evaluation of a computer-assisted remedial algebra curriculum on attitudes and performance of ninth-grade English learners* (Doctoral dissertation). Retreived from ProQuest Dissertations and Theses database. (ISBN No. 978-1-109-29165-0).

Watts, M., & Ebbutt, D. (1987). More than the sum of the parts: Research methods in group interviewing. *British Educational Research Journal, 13*, 25–34. doi: 10.1080/0141192870130103

Wayman, J. C., Cho, V., & Johnston, M. T. (2007). *The data-informed district: A district-wide evaluation of data use in the Natrona county School District.* Austin: University of Texas.

Wylie, E. C. (2008). *Formative assessment: Examples of practice.* Washington, DC: Council of Chief State School Officers. Retrieved from: http://www.ccsso.org/Resources/Programs/Formative_Assessment_for_Students_and_Teachers_(FAST).html

Yeo, S. (2010). Predicting performance on state achievement tests using curriculum-based measurement in reading: A multilevel meta-analysis. *Remedial and Special Education, 31*, 412–422. doi: 10.1177/0741932508327463

Young, V. M. (2008). Supporting teachers' use of data: The role of organization and policy. In E. B. Mandinach & M. Honey (Eds.), *Linking data and learning* (pp. 87–106). New York: Teachers College Press.

Young, V. M., & Kim, D. H. (2010). Using assessments for instructional improvement: A literature review. *Education Policy Analysis Archives, 18*(19), 1–40.