The Accuracy of U.S. Middle School Teachers' Judgment of Student Reading Abilities

Deborah K. Reed,[1] Adam Reeger,[2] Eliot Hazeltine,[3] and Bob McMurray[3]
[1]Tennessee Reading Research Center, University of Tennessee
[2]Alpine Testing Solutions
[3]University of Iowa

Abstract

Teacher judgments of students' reading abilities in the elementary grades have been researched extensively, but less is known about how middle school teachers judge their students' word reading, fluency, vocabulary, and comprehension skills. Such information could be useful when determining which students and reading components would be reasonable instructional priorities. Thus, the present study explored U.S. teachers' accuracy at predicting the performance of students in Grades 6–8 on standardized measures of reading abilities. The multilevel analytic models accounted for the nesting of students ($n = 97$) within teacher raters ($n = 12$) at three middle schools in one school district. Results indicated that the teachers' ratings of overall ability and their beliefs about the specific skills with which their students struggled were poor predictors of actual student performance. Although the small sample of teachers from one district in one country limits the generalizability of the results, the findings suggest that some middle school teachers' judgments could misidentify students at potential risk for reading difficulties or misalign instruction with particular skill areas of need. The discussion addresses the importance of making efficient but accurate decisions about screening students for reading intervention and planning differentiated or targeted intervention.

Keywords: Reading ability, middle school, teacher judgment

Adolescents demonstrate a wide range of reading abilities (Firmender et al., 2013). Even those not performing proficiently make up a heterogeneous group, with different profiles of skills and often requiring specialized instruction in multiple components (Foorman et al., 2017; Oslund et al., 2018). This broad range of abilities presents challenges to middle school teachers as they attempt to prioritize (a) who among their students are most in need of extra support and (b) what to address in the limited instructional time available (Jaeger & Pearson, 2017). In U.S. middle schools, the previous year's score on state-mandated summative assessments has been considered a strong and efficient predictor of an adolescent's performance in the current year (Fuchs et al., 2010; Nelson et al., 2016; Stevenson, 2017). Nevertheless, these measures of grade-level reading achievement were not designed to indicate the particular reading skills (e.g., word reading, fluency, vocabulary, comprehension) that underly each student's difficulty (O'Reilly et al., 2012).

It has been proposed that teacher judgments may improve the screening of adolescents for instructional planning purposes (Nelson et al., 2016), but little is known about how middle school teachers judge the reasons why students might be experiencing reading

Reed, D. K., Reeger, A., Hazeltine, E., & McMurray, B. (2023). The Accuracy of U.S. Middle School Teachers' Judgment of Student Reading Abilities. *International Journal for Research in Learning Disabilities, 6*(1), 14-27. https://doi.org/10.28987/ijrld.6.1.14

difficulty. Therefore, the present study explored the extent to which teachers were accurate predictors of middle school students' word reading, fluency, vocabulary, and comprehension abilities.

Teacher Judgments of Student Abilities

It can be both time- and resource-intensive to collect detailed assessment data on students' reading abilities. By contrast, teacher data can be more easily collected, in part, because teachers observe students' reading behaviors on a daily basis and naturally form impressions about their strengths and weaknesses (Kettler & Albers, 2013; Speece et al., 2011). Thus, teacher judgments are considered more efficient and less costly than administering multiple tests to students (Speece et al., 2010).

Yet, studies show that not all teacher judgments take the same form. For example, studies conducted in the United States, Canada, and Austria had teachers rate students' reading ability on a scale (e.g., Bailey & Drummond, 2006; Beswick et al., 2005; Feinberg & Shapiro, 2009; Nelson et al., 2016; Paleczek et al., 2017; Valdez, 2013). Alternatively, teachers in Australia rank-ordered students by perceived ability or percentile rank of performance (e.g., Bates & Nettelbeck, 2001; Madelaine & Wheldall, 2005). In some cases, U.S. teachers predicted specific scores on a reading test (e.g., Feinberg & Shapiro, 2003; Martin & Shapiro, 2011), but in other studies teachers judged students broadly by status such as proficient or not, at risk or not, high or low achieving, and word caller or comprehender (e.g., Hamilton & Shinn, 2003; Nelson et al., 2016). These different judgment types have been used alone or in combination (e.g., Begeny et al., 2011; Missall et al., 2019; Speece et al., 2010, 2011).

Although not focused on reading exclusively, a meta-analysis examining 75 international studies of the relation between teacher judgments and students' academic achievement found an overall mean effect of 0.63 (Südkamp et al., 2012). The concordance was not moderated by judgment type (rating, ranking, or score prediction) or by the number of points on a rating scale. However, the correlation between teachers' judgment and students' achievement was higher when teachers were informed about the test against which their judgments were to be compared (as opposed to an uninformed judgment) and when the teachers were rating a specific domain that matched a tested domain. The meta-analysis included research from kindergarten through twelfth grade, but Südkamp et al. did not analyze whether grade level was associated with concordance. When judging students' reading abilities, grade level may be important because middle school teachers are often not well trained in teaching the component skills of reading like decoding, fluency, or comprehension (Heller & Greenleaf, 2007; Kosanovich et al., 2010).

Previous Investigations of Judging Students' Reading Abilities

Most of the extant literature on teacher judgments of students' reading abilities has been conducted in the elementary grades in the United States, Canada, and Australia. Results have been inconsistent. Occasionally, studies report that teachers were more likely to underestimate students' abilities (Beswick et al., 2005), but more often studies report that teachers overestimated students' performance (e.g., Bates & Nettelbeck, 2001; Hamilton & Shinn, 2003; McKevett & Kiss, 2019). Further, some have argued that teachers are fairly accurate (Feinberg & Shapiro, 2009; Kettler & Albers, 2013; Missall et al., 2019), while others have raised concern about the number of students misjudged – particularly at lower ability levels (e.g., Begeny et al., 2011; Madelaine & Wheldall, 2005; Martin & Shapiro, 2011).

Several researchers recommended including teacher judgments with other test scores to classify elementary students with or without risk of reading difficulties (Kettler & Albers, 2013; Missall et al., 2019; Speece et al., 2010, 2011). A similar suggestion was made in a study that examined U.S. participants in Grades 6 through 8 exclusively (Nelson et al., 2016). When included in a regression model with teacher ratings, the previous year's summative reading assessment was still the strongest single predictor of student outcomes in the current year. Nonetheless, Nelson et al. proposed that adding teacher judgments might improve the efficiency of screening students for intervention while reducing the resource burden on schools. However, the middle school teachers were not asked to identify the particular reading skills with which they thought their students struggled.

Contributions of the Present Study

In addition to focusing on Grades 6–8, the present study sought to address several gaps in the existing literature on teacher judgments of students' reading abilities. First, few studies have assessed multiple domains of reading using standardized tests (Bailey & Drummond, 2006; Feinberg & Shapiro, 2009; Speece et al., 2010, 2011). Most often, teacher judgments have been explored in relation to oral reading fluency screening instruments (e.g., Begeny et al., 2011; Martin & Shapiro, 2011; McKevett & Kiss, 2019; Missall et al., 2019) and criterion-referenced assessments, including annual summative assessments (e.g., Begeny et al., 2011; Kettler & Albers, 2013; Missall et al., 2019; Nelson et al., 2016).

Second, previous researchers have rarely accounted for the nesting of students within teacher raters. One identified study used multilevel modeling to evaluate teachers' ratings of the decoding and comprehension abilities of their second- and third-grade students in Austria (Paleczek et al., 2017). Raters likely are an important source of unexplained variance, and students are not randomly assigned to teachers (e.g., some teacher may have more high or low performers than would be expected by chance). Therefore, analyses that take into account the nested nature of the data are important to improving our understanding the predictive utility and consequential validity of teacher judgments.

Finally, prior studies of teachers' accuracy at judging student performance typically have been limited to rating status such as at risk or proficient (e.g., Feinberg & Shapiro, 2009; Kettler & Albers, 2013; Madelaine & Wheldall, 2005; McKevett & Kiss, 2019; Missall et al., 2019). That is, researchers have analyzed group status indicators and not continuous scores (Speece et al., 2010, 2011). This practice has been highlighted as obscuring the inaccuracies of identifying students within the large and typically dichotomous groups (Bates & Nettelbeck, 2001; Begeny et al., 2011; Feinberg & Shapiro, 2003). Focusing only on broad groupings may overestimate the accuracy of teachers' judgments or fail to detect whether teacher ratings were equally predictive for children at all ability levels. Hence, there is a need for research that adopts continuous metrics.

Purpose and Research Question

Assessing students' reading abilities can be a time-intensive process, and middle school teachers do not always agree with the resulting data nor use the data to make their instructional decisions (Reed, 2015; Deeney & Shim, 2016). Therefore, it is important to understand how teachers' judgments of their students' abilities align with the students' performance on reading tests of those abilities in the middle grades. To that end, this study was designed to answer the primary research question: How well did teacher ratings predict the performance of U.S. students in Grades 6–8 on objective measures of students' reading abilities?

Educator judgments could be particularly important when planning differentiated instruction or targeted interventions for students who may be struggling with one or more reading skills (Fien et al., 2018; Oslund et al., 2018). It could be teachers are better at judging certain reading skills more so than others (Bailey & Drummond, 2006; Hamilton & Shinn, 2003; Paleczek et al., 2017). Thus, for the subgroup of struggling readers, the exploratory research question asked if U.S. teachers were differentially capable of assessing several skills, including decoding, vocabulary, fluency, and comprehension. Finally, it could be teachers are better at rating students at certain proficiency levels (Begeny et al., 2011; Feinberg & Shapiro, 2009; Madeleine & Wheldall, 2005; McKevett & Kiss, 2019), which would have implications for the use of teacher ratings in planning differentiated instruction or targeted interventions. Thus, the other exploratory research question was: How did teachers' ratings predict the reading skill performance of U.S. students in Grades 6–8 at different locations along the achievement continuum?

Method

Setting and Participants

The study involved data collected during the start of the second semester at three middle schools from a midsize city in a U.S. Midwestern state. At the time of the study, the only routine reading assessment administered to students was the annual state-mandated summative assessment of reading achievement. Approximately 36% of students in the participating middle schools had not achieved the grade-level proficiency benchmark on their previous year's summative assessment, and another 28% were within the confidence interval. Taken together, about 64% of all students demonstrated potential risk of not reading proficiently in the year of the study. Participants included both the teachers who completed the ratings and their students who were rated.

Teachers

A total of 12 teachers (School A = 4 teachers; School B = 5; School C = 3) consented to participate and provided ratings of their students' performance. The schools separated their literacy and language arts instruction into two different class periods. Four participants were designated as literacy teachers (one of whom also taught special education), and five participants were designated as language arts teachers (one of whom also served part-time as an instructional coach). There were two teachers who taught both literacy and language arts to English learners. The final teacher in the study taught only students in special education. As can be seen in Table 1, about

Table 1
Number of Students Rated Within Each Teacher Role Classification

	Teacher Role				Grade-Level Total
Grade	Literacy	ELA-ELL	Special Ed	LA	
6	29	2	1	19	51
7	11		4	9	24
8	9	11		2	22
Total	49	13	5	30	97

Note. ELA-ELL = literacy and language arts teacher for students who were English learners; LA = language arts instructor.

half the students ($n = 49$) were rated by literacy teachers. All teachers were involved in the district's multitiered system of supports model, which required them to help identify students not responding to instruction who might need differentiated instruction or targeted supplemental intervention.

Students

To be eligible for the study, students had to be enrolled in Grades 6–8 and have scored between the 10th and 60th percentiles on the previous year's state reading test. Thus, they were considered average to below-average readers who might be considered at low to high risk of not meeting grade-level reading expectations. A total of 97 eligible students in the 12 participating teachers' classes had parental consent and granted assent (School A = 41 students; School B = 28; School C = 28). Students subsequently completed the four measures of reading ability used in the analyses (see section on measures). Most of the student participants were receiving free or reduced-price lunch ($n = 75$), a proxy for economic disadvantage; just over half were female ($n = 56$); and less than a quarter ($n = 22$) were receiving special education services. As shown in Table 1, each student was rated by one teacher, and the greatest number of ratings in the dataset were made on sixth-grade students ($n = 51$).

Measures

Students

Trained researchers administered tests of word reading, fluency, vocabulary, and comprehension to all students. Given the number of measures, administrations were counterbalanced and distributed across five testing days such that students did not test for more than an hour each day. In the following week, make-up testing was conducted with any students who were absent the previous week.

All testing occurred in quiet rooms at the school that were not being used for instruction at the time. Assigned testers were responsible for the initial scoring of the measures they administered, and they were monitored by the research coordinator throughout for fidelity to the testing protocols. After all testing was complete, test documents were checked by an independent rater for accuracy and completeness.

Word Reading. Two untimed subtests of the Woodcock Johnson Reading Mastery Test (WRMT; reliability = .91 to .97; Woodcock et al., 2001) were individually administered to students. For Word Identification, scores were based on the number of isolated English words students accurately read aloud. The words progressed from those that were high frequency to words of increasing difficulty. In the Word Attack subtest, students were scored on the number of pseudowords they were able to decode with phonetic accuracy. For each subtest, students first completed sample items and then proceeded until they made four consecutive errors or completed the final item. Raw scores on both subtests were converted to standard scores for use in the analyses. Because the measures assessed sight word reading as well as decoding, this domain is collectively referred to as word reading.

Fluency. Oral reading fluency rate was individually assessed with the Texas Middle School Fluency Assessment (TMSFA; Francis et al., 2010), for which students read a series of three passages aloud for 1 min each while the examiner recorded words that were mispronounced, skipped, substituted, or provided by the examiner after a 3 sec hesitation. Raw scores for each passage were calculated as the number of words read correctly in the minute, and those were converted into equated scores that accounted for passage difficulty. Data used in the analyses were the average of the three equated scores. In a previous confirmatory factor analysis with data from students in Grades 7 and 8, the factor loading of TMSFA average equated scores on a fluency construct was .901 ($p < .001$; SE = .019; residual variance = .187; Reed et al., 2012).

Vocabulary. Oral vocabulary knowledge was assessed using the Peabody Picture Vocabulary Test (PPVT; reliability = .93 to .94; Dunn & Dunn, 2007). Students had to choose from four picture options the image that represented each word the examiner said. After the training items, testing began with age-based starting sets, but these were adjusted downward until students met the basal requirement of making no more than one error in a set. Testing was discontinued when students made eight or more errors in a set. Raw scores were then converted to standard scores. The individually administered test was untimed, and scores were based on the number of correctly identified items.

Written vocabulary knowledge was assessed with the group-administered Vocabulary subtest of the Gates-MacGinitie Reading Test (GMRT; reliability = .90 to .92; MacGinitie et al., 2000). Students had 20 min to choose among multiple-choice options the words or phrases that were similar in meaning to underlined words provided in brief contexts. The raw number of items correct was converted to a scale score because the developer did not provide a standard score conversion.

Comprehension. Silent reading comprehension was assessed with the group-administered Comprehension subtest of the GMRT (reliability = .91 to .92; MacGinitie et al., 2000). Students had 35 min to read a series of short passages and answer multiple-choice questions associated with each passage. The raw number of items correct was converted to a scale score because the developer did not provide a standard score conversion.

Teachers

Reader Rating Form. Without knowledge of the tests administered in this study or test scores, teachers were asked to rate their students' overall abilities and, for students judged to be below grade level, their performance status in the same four domains as tested with objective measures. Teachers were aware that the project concerned identifying students' specific reading abilities to better understand their overall performance and to inform the design and delivery of reading instruction. Given their lack of familiarity with the study measures, teachers were considered uninformed about those tests but informed of the state assessment qualifying students for participation in the project (Südkamp et al., 2012).

Using an instrument developed by Speece et al. (2011), participating teachers rated each of their participating students on Overall Reading ability using a 5-point Likert scale (1 = far below grade level; 2 = below grade level; 3 = on grade level; 4 = above grade level; 5 = far above grade level). The subgroup of students rated with a 1 or 2 was further dichotomously rated to indicate what the teacher believed to be contributing to each student's lack of proficiency (Word Reading, Fluency, Vocabulary, Comprehension). The teacher ratings of Motivation were removed from the Speece et al. (2011) form to focus only on the reading skills that could be objectively measured. For each skill, teachers indicated a student's status as "0" if perceived not to have a difficulty or "1" if perceived to have a difficulty in the specific area. Teachers could indicate one area or a combination of areas. Of the 97 student participants, 45 were rated for a difficulty with the four reading skills.

The overall rating and the dichotomously identified areas of difficulty were used in the analyses. To make the data directionally similar to the Likert-scale ratings and to facilitate correct interpretation of the correlations and slopes, the 0/1 indicators were reverse-coded before the data were analyzed. It was confirmed that this reverse coding did not change the magnitude of the slopes but simply the sign or direction of the relationship (i.e., positive, as opposed to negative, slopes would indicate agreement between teacher judgments and scores).

Analytic Approach

In order to compare how well teacher ratings predicted student performance on each of the reading measures, linear multilevel mixed-effects models were fit separately for each combination of reading measure and teacher rating category using the lme4 package in R (Bates et al., 2015; R Core Team, 2017). These models take into account the hierarchical structure of the data; namely, that students were nested within teachers who rated them. Furthermore, these models incorporate both fixed effects that are effects measured as constant across individuals and random effects, which are effects that are allowed to vary across individuals (Raudenbush & Bryk, 2002). The only fixed effect in each model was the continuous teacher rating score. The analyses additionally incorporated a random intercept effect for teacher. This means the mean reading measure scores of students were allowed to vary across teachers.

The model may be expressed as:

Level–1

$$Y_{ij} = \pi_{0j} + \pi_{1j} X_{ij} + e_{ij}$$

Level–2

$$\pi_{0j} = \theta_0 + b_{0j}$$

where Y_{ij} is the score on reading measure Y for the i^{th}

student (rated by the j^{th} teacher), and X_{ij} is the rating given for the i^{th} student by the j^{th} teacher for reading category X. At the first level of the model, π_{0j} is the mean reading measure score for students rated by teacher j, π_{1j} is the regression slope relating ratings on X from teacher j to reading measure scores, and the residual term for student is assumed to be Gaussian $e_{ij} \sim N(0, \sigma^2)$. At the second level of the model, θ_0 is the grand mean score on the reading measure Y, and $b_{0j} \sim N(0,\tau_0)$ is the random intercept for teacher j. This specification of the model accounts for two residual variances: the between-teachers' variability (Level 2) and the within-teachers' variability (Level 1). For each model, every student was assigned a rating and had a full set of reading measure scores, so there were no missing data concerns.

Because teacher ratings of student reading ability may predict student performance on reading measures differently at different levels of reading performance, separate linear quantile mixed-effects models were fit for each combination of reading measure and teacher rating category, thus allowing for comparisons of effects at different locations along the achievement continuum (Koenker & Hallock, 2001). The model may be conceptualized as follows: Let a sample of observations (x_i^T, y_i) be drawn from a population with continuous distribution function $F_{y_i|x_i}$. The quantile function is defined as its inverse, $Q_{y_i|x_i} = F^{-1}{}_{y_i|x_i}$, and in the linear case, $Q_{y_i|x_i}(\boldsymbol{\tau}) = \mathbf{x}_i^T\beta^\tau$, where $\boldsymbol{\tau}$ is the quantile level of interest, ranging from 0 to 1. The $\boldsymbol{\tau}^{th}$ linear regression quantile is defined as the solution of

$$\min_{\beta \in Rp} \sum \rho_\tau (y_i - x_i^T \beta),$$

where ρ_τ is the asymmetrically weighted L_1 loss function (Geraci, 2014). These models were fit using the lqmm package in R (Geraci, 2018) and similarly accounted for nesting.

Results

Descriptive statistics for students' scores on the reading assessments and the correlations among the scores may be found in Tables 2 and 3, respectively. Note in Table 2 that the two subtests on the WRMT had very similar means, as did the two subtests on the GMRT. In addition, subtest scores within an assessment were more highly correlated than scores from different assessments (see Table 3). WRMT Word Identification and the TMSFA were each significantly correlated with performance on all the other measures. WRMT Word Attack generally had lower correlations with students' performance on the other measures, except for the TMSFA. Overall, the correlation values suggest that the tests were measuring independent constructs as anticipated. Similar descriptive statistics and tetrachoric correlations for the teacher ratings of student reading ability are provided in Tables 4 and 5, respectively. Statistically significant correlations were found for Vocabulary with Word Reading and for Vocabulary with Fluency, but ratings for Comprehension showed little relation with ratings in any of the other reading categories.

Association Between Teacher Ratings and Student Test Performance

Table 6 shows the associations between teacher ratings of student reading ability and student performance on each of the reading measures, as represented by standardized regression slopes obtained from the multilevel mixed-effects models (one for each combination of reading measure and teacher rating category). Recall that the models take into account the variability among raters. Each combination of reading measure and teacher rating category was fit as a separate model, resulting in 30 total models. The first column of Table 6 provides standardized slopes relating teacher ratings of students' Overall Reading ability to student scores on each of the reading measures. Although significance testing is reported for each slope to help establish a baseline for what may be meaningful, the high number of comparisons makes it unwise to draw strong conclusions about any individual test. Thus, the focus primarily is on the relative effect sizes (represented by the standardized coefficients) and the patterns across measures and domains.

Results show that Overall Reading ratings most strongly predicted scores on the TMSFA, followed by moderate predictions of student performance on GMRT Vocabulary and WRMT Word Identification. Scores on WRMT Word Attack and GMRT Comprehension were more weakly predicted by ratings of Overall Reading, and students' performance on the PPVT had almost no association with teachers' overall rating.

The remaining columns of Table 6 show how teacher ratings within specific reading categories (Word Reading, Vocabulary, Fluency, and Comprehension) predicted scores on each of the specific reading measures. This was only done for the subsample of students who teachers believed were performing far/below grade level, roughly half of the full sample. Generally, these slopes were small, and some

Table 2
Descriptive Statistics for Reading Assessment Measures (N = 97)

Measure	*Mean*	*SD*	Min	Max
WRMT Word Identification	89.3	13.3	56	123
WRMT Word Attack	90.8	18.4	0	122
GMRT Vocabulary	505.1	24.4	452	556
GMRT Comprehension	505.6	27.9	398	563
TMSFA	125.9	34.0	40	214.3
PPVT	95.4	9.9	71	117

Note. WRMT = Woodcock Reading Mastery Test; GMRT = Gates-MacGinitie Reading Test; TMSFA = Texas Middle School Fluency Assessments; PPVT = Peabody Picture Vocabulary.

Table 3
Correlations Among Reading Assessments

Measure	WRMT-WID	WRMT-WA	GMRT Vocab	GMRT Comp	TMSFA	PPVT
WRMT-WID	1	0.588**	0.491**	0.454**	0.543**	0.220*
WRMT-WA		1	0.244*	0.186	0.417**	0.193
GMRT Vocab			1	0.702**	0.533**	0.477**
GMRT Comp				1	0.555**	0.335**
TMSFA					1	0.060
PPVT						1

Note. WRMT = Woodcock Reading Mastery Test; WID = Word Identification subtest; WA = Word Attack subtest; GMRT = Gates-MacGinitie Reading Test; Vocab = vocabulary; Comp = comprehension; TMSFA = Texas Middle School Fluency Assessments; PPVT = Peabody Picture Vocabulary Test.

**p* < .05, ***p* < .01.

Table 4
Descriptive Statistics for Teacher Rating Categories

		Number of Ratings in Category					
Category	*Mean*	0	1	2	3	4	5
Overall Reading[a]	2.51	--	15	30	41	10	1
Word Reading[b]	0.53	21	24	--	--	--	--
Vocabulary[b]	0.78	10	35	--	--	--	--
Fluency[b]	0.64	16	29	--	--	--	--
Comprehension[b]	0.80	9	36	--	--	--	--

Note. [a]Overall reading ability was rated on a scale of 1–5 for all 97 participants. [b]This category was dichotomously rated by teachers (0 = not an area of difficulty; 1 = area of difficulty) only for the 45 students whom teachers rated with a 1 or 2 for overall reading ability.

Table 5
Correlations Among Teacher Rating Categories (N = 45)

Category	Word Reading	Vocabulary	Fluency	Comprehension
Word Reading	1	0.598*	0.370	0.159
Vocabulary		1	0.610*	0.246
Fluency			1	0.166
Comprehension				1

**p* < .05.

Table 6
Standardized Slopes from Multilevel Models Relating Teacher Ratings to Scores on Reading Assessments

	Teacher Ratings				
Measure	Overall Reading *(N = 97)*	Word Reading *(n = 45)*	Vocabulary *(n = 45)*	Fluency *(n = 45)*	Comprehension *(n = 45)*
WRMT-WID	0.308**	0.198	0.174	0.327	0.008
WRMT-WA	0.204*	0.078	-0.060	0.191	0.088
GMRT Vocab	0.343**	0.012	0.306	0.267	0.098
GMRT Comp	0.205*	0.126	0.432**	0.104	0.157
TMSFA	0.482**	0.141	0.218	0.123	-0.010
PPVT	0.095	-0.135	-0.095	-0.230	0.038

Note. WRMT = Woodcock Reading Mastery Test; WID = Word Identification subtest; WA = Word Attack subtest; GMRT = Gates-MacGinitie Reading Test; Vocab = vocabulary subtest; Comp = comprehension subtest; TMSFA = Texas Middle School Fluency Assessments; PPVT = Peabody Picture Vocabulary Test.

**p* < .05, ***p* < .01.

ratings negatively predicted student performance on reading measures (e.g., Fluency had a small negative relation with PPVT). Only the Vocabulary rating showed a moderate positive association with the GMRT Vocabulary and Comprehension subtests, and only its association with GMRT Comprehension was statistically significant (at the $\alpha = 0.01$ level).

Vocabulary also was the only rating for which teachers demonstrated moderate accuracy in judging a particular area of student difficulty. However, this was only true with respect to written vocabulary (GMRT Vocabulary slope = 0.306), not oral vocabulary (PPVT slope = -0.095). The other teacher category ratings were weak predictors of students' performance on assessments designed to measure the same reading ability. Specifically, participating teachers were poor judges of whether word reading, fluency, or comprehension were areas of difficulty for their middle school students perceived to be reading far/below grade level. Teacher ratings for fluency were more strongly associated with WRMT Word Identification than the TMSFA, and their ratings for vocabulary were more strongly associated with the TMSFA than either the PPVT or the GMRT Vocabulary.

Teacher Ratings by Student Ability

To understand how the relation between teacher ratings of student reading ability and student performance on reading measures may differ at different levels of test performance, we turn to the results of exploratory linear quantile mixed-effects models. Separate quantile models were fit for each combination of reading measure and teacher rating category, resulting in 30 models that estimated parameters at the 25th, 50th, and 75th percentile values of reading measure performance. The standardized slopes from each of these models are provided together in one table (Table 7).

Predictions near the extreme percentiles of reading performance (10th or 90th) were not considered because issues of bias or measurement error may become exacerbated with small sample sizes at the

Table 7
Standardized Slopes From Quantile Regression Models Relating Teacher Ratings to Reading Assessment Scores

	Teacher Ratings				
Measure	Overall Reading *(N = 97)*	Word Reading *(n = 45)*	Vocabulary *(n = 45)*	Fluency *(n = 45)*	Comprehension *(n = 45)*
			25th Percentile		
WRMT-WID	0.412	0.094	0.184	0.367	< |0.001|[a]
WRMT-WA	0.140	< |0.001|a	-0.109	0.094	0.183
GMRT Vocab	0.413**	-0.011	0.405	0.114	0.095
GMRT Comp	0.177*	0.329	0.308*	-0.296	< |0.001|[a]
TMSFA	0.446**	0.298	0.242	0.231	-0.179
PPVT	0.190	< |0.001|[a]	< |0.001|[a]	-0.254	0.194
			50th Percentile		
WRMT-WID	0.309	0.159	0.212	0.214	-0.026
WRMT-WA	0.250	-0.009	-0.152	0.188	0.131
GMRT Vocab	0.375**	0.055	0.249	0.339	0.168
GMRT Comp	0.262	0.269	0.291*	0.125	0.025
TMSFA	0.421**	0.211	0.242	0.160	0.037
PPVT	0.184	-0.080	-0.079	-0.285	0.155
			75th Percentile		
WRMT-WID	< |0.001| [a]	0.191	0.133	0.336	-0.153
WRMT-WA	0.398*	0.228	-0.234	0.250	-0.137
GMRT Vocab	0.388**	0.088	0.294	0.275	0.106
GMRT Comp	0.295	0.091	0.466**	0.359	0.262
TMSFA	0.462**	0.054	0.205	-0.005	0.028
PPVT	-0.046	-0.339	0.081	0.002	0.080

Note. [a]Standardized slopes that are less than 0.001 in absolute value terms are notated <|0.001|; WRMT = Woodcock Reading Mastery Test; WID = Word Identification subtest; WA = Word Attack subtest; GMRT = Gates-MacGinitie Reading Test; Vocab = vocabulary subtest; Comp = comprehension subtest; TMSFA = Texas Middle School Fluency Assessments; PPVT = Peabody Picture Vocabulary Test.

**p* < .05, ***p* < .01.

extremes of score distributions (Akram et al., 2013; Lockwood & Castellano, 2016). In the current study, the sample divided into percentiles for measuring the effect of Overall Reading ratings (*n* = 97) was different from the sample divided into percentiles for measuring the prediction of the specific category ratings because only those students whom teachers rated as reading far/below grade level (*n* = 45) were further rated by area of difficulty. This means that for models that only included students rated as reading far/below grade level, the percentiles of performance would be much lower than the corresponding percentiles of performance for models that included the whole student sample (e.g., the 75th percentile for the far/below-level group represents a much lower level of performance than the 75th percentile for the whole group). Table 7 reports the estimated slopes at the three designated percentiles from all quantile models.

As illustrated, the Overall Reading rating was found to be a statistically significant predictor of GMRT Vocabulary and TMSFA scores at all three quantiles, with the associations demonstrating similar magnitudes. However, there were distinct differences across the quantiles for many of the slopes relating teacher rating categories to performance on the reading measures, and only the relation of Vocabulary and GMRT Comprehension consistently demonstrated statistical significance across quantiles—though still with different magnitudes of associations and none that were particularly robust. Teacher ratings tended to predict reading performance more strongly at some quantiles of performance than at others, but there were no discernible patterns found in these differences across quantiles.

Some models showed stronger associations between rating and reading measure at the 25th percentile of performance. For example, Vocabulary ratings moderately predicted GMRT Vocabulary, and both Vocabulary and Word Reading ratings moderately predicted GMRT Comprehension scores. Nevertheless, only the relation between Vocabulary rating and GMRT Comprehension was statistically significant at all three quantiles. Other models showed stronger associations at the 50th percentile of performance: Fluency ratings moderately predicted GMRT Vocabulary, but the relation was not statistically significant. Finally, other associations were strongest at the 75th percentile, though none were significant. For example, Fluency ratings moderately predicted WRMT Word Identification and GMRT Comprehension, and Word Reading ratings had a moderate negative association with PPVT scores.

Additionally, although there were differences in prediction across quantiles within every teacher rating category, Vocabulary had the most consistent relations with reading measures (i.e., the smallest differences in slopes) across quantiles. Fluency ratings, on the other hand, had the least consistent associations with reading measures across performance quantiles. In summary, teacher ratings of student reading ability predicted student reading performance differently at different levels of student performance, but the direction and magnitude of these differences depended on the reading measure and rating category considered and rarely were significant.

As with the results of the other multilevel models, the quantile models revealed that teachers' beliefs about the specific areas of reading with which their students struggled were poor predictors of students' actual performance on the reading measures, regardless of students' level of performance on the associated reading measures. That is, rating a student with a weakness in word reading did not predict the student's word reading assessment performance at the 25th, 50th, or 75th percentile; rating a weakness in fluency did not predict fluency performance at any percentile; and so on. The only exception was found in teachers' moderately accurate judgments of whether the subset of students performing in the 25th percentile had a specific difficulty with written vocabulary knowledge (GMRT Vocabulary slope = 0.405).

Discussion

Instruction for adolescents with heterogeneous reading profiles should be informed by data that map the difficulties each student is experiencing (Fien et al., 2018; Jaeger & Pearson, 2017; Oslund et al., 2018). However, such data may be rare in U.S. middle schools, as typical state-mandated summative assessments measure only grade-level reading achievement (O'Reilly et al., 2012). To remedy this situation, it has been suggested that teachers might provide an efficient data source to complement annual summative assessments in screening adolescents for risk of not reading proficiently (Nelson et al., 2016). To inform the use of data at the middle-school level, this study sought to determine the association between objective test scores and U.S. teachers' judgments about their students' reading performance.

The first research question asked whether participating teachers' ratings predicted overall reading ability. Unlike prior studies, the approach in the present study accounted for the nesting of students within raters and found that teacher ratings for students' Overall Reading ability (i.e., far below, below, on, above, or far above grade level) were not strong predictors of the reading skill scores of students who had performed between the 10th and 60th percentiles on the previous year's summative assessment. Teacher judgments were moderately related to fluency rate (0.482), written vocabulary (0.343), and word identification scores (0.308) but weakly related to passage comprehension (0.205) and decoding scores (0.204). Moreover, Overall Reading ratings had no relation to students' oral vocabulary, even though the assessment used to measure this skill (PPVT) was moderately and significantly correlated with written vocabulary knowledge (0.477) and comprehension (0.335).

In general, the degree of inaccuracy among participating teachers would suggest little would be gained by querying their judgments about the overall performance of students already screened for risk with the annual criterion-referenced assessment. Moreover, teachers were only moderately successful at predicting some domains (vocabulary) but not others (fluency). The prediction of teacher ratings also differed even within closely related domains (word identification but not decoding, oral but not written vocabulary). These findings suggest that the teachers' overall estimations of their student abilities may have been idiosyncratically biased to certain skills or behaviors and may have missed deficits in other crucial areas such as comprehension. Hence, results of the present study echo the cautions about misidentification raised by researchers of teacher judgments in elementary school (e.g., Begeny et al., 2011; Madelaine & Wheldall, 2005; Martin & Shapiro, 2011).

Because specialized instruction in some reading components might be planned only for students ultimately determined to be at greater risk (Fien et al., 2018), the second research question was addressed by analyzing teacher-provided status indicators (i.e., exhibiting a difficulty or not) for the subgroup of students that the teachers rated as reading far/below grade level. Although a prior meta-analysis found teacher ratings were congruent with domain-specific measures (Südkamp et al., 2012), results of the exploratory inquiry in the present study suggested that teacher judgments of the specific sources of students' reading difficulties bore almost no relation to students' actual scores on the corresponding test(s) of those skills. The only exception was that teachers were moderately accurate in indicating which students did/not have difficulty in the area of written vocabulary (0.306). However, in addressing the final research question, the exploratory quantile regression revealed that this moderate association between teacher judgment and student vocabulary performance only held true for students performing at the lowest levels (0.405). This suggested the teachers were not truly more accurate at judging written vocabulary knowledge.

Across all other skills and quantiles analyzed, teacher judgments had weak, no, or negative relations to the corresponding test scores for those specific skills. There was a particularly striking lack of concordance for comprehension. Whereas Paleczek et al. (2017) suggested elementary teachers were better at identifying comprehension than decoding skills, middle school teachers' judgments of which students did/not have comprehension difficulties in the present study were not related to judgments in any other category or to any of the assessment scores across quantiles of student performance. Other researchers have noted comprehension is a complex, multifaceted construct that is challenging to measure (Betjemann et al., 2011; Eason et al., 2012). Yet, despite the incomplete picture that the GMRT Comprehension subtest might present of students, scores were mostly significantly and moderately correlated to students' scores on tests of other skills. In contrast to the picture afforded by the objective measures, comprehension disassociated from the other skills in teacher ratings.

It should be noted that U.S. middle school teachers face enormous challenges in planning literacy instruction that meets a variety of individual needs and simultaneously prepares students for rigorous college and career readiness standards (Jaeger & Pearson, 2017). Moreover, it is not common for U.S. middle and high school educators to have received adequate preservice training in reading development because their teacher preparation programs focus on content-area demands such as literary appreciation (Heller & Greenleaf, 2007; Kosanovich et al., 2010).

The results of this study should not be used to blame teachers for a lack of knowledge. Importantly, it was not possible to determine from the data why teacher perceptions were so discordant because teachers were not asked how they were defining the literacy skills or on what basis they were judging students' abilities. The exploratory study was designed to inform the potential consequential validity of using the judgments of typical U.S. middle school teachers to identify the students and skills that might be prioritized for specialized instruction.

Limitations and Directions for Future Research

As defined by Südkamp et al. (2012), teachers were uninformed about the tests of specific reading skills that were administered for this project. Even though the teachers knew that the project concerned identifying students' specific reading abilities, they only knew about the annual summative assessment and students' performance on that criterion-based measure. It is possible that teachers simply did not share the same definitions of word reading, fluency, vocabulary, and comprehension that were applied in the measures we administered. To better understand teachers' judgments, future research in the middle grades may take an approach similar to that of Bailey and Drummond (2006), who asked elementary teachers to rationalize their ratings and provided them with guidelines to improve the concordance of their ratings with the test scores.

In addition, the present study included a relatively small number of teachers ($n = 12$), only one of whom was a special educator. This precluded analyzing differences in teachers' judgment accuracy by instructor assignment or expertise. The data largely were concentrated in Grade 6 ($n = 51$, 53%), and the status judgments on specific reading skills were obtained on only the smaller subsample of students analyzed for specific areas of reading difficulty ($n = 45$). The small sample rendered the quantile regression underpowered and the estimation of random effects imprecise. Relatedly, the study was limited to one school district in one country, so the results are not necessarily generalizable to the broader population of middle school teachers in other U.S. states or other countries. Thus, the results of this study offer an exploratory look at middle school teacher judgments. When combined with the inconsistent results found across the numerous studies of teacher judgments in the elementary grades, it would be premature to draw strong conclusions from these findings. Additional research is needed to enrich the literature base for Grades 6–8.

Implications

The results not only suggest that some teachers' judgments may misidentify students with potential risk for reading difficulties, but also that teachers' perceptions of the skills contributing to poor performance may lead to planning instruction that is not aligned to students' actual needs. This match between teacher judgments and subsequent intervention planning would be critical to better equipping students for reading success (Fien et al., 2018; Jaeger & Pearson, 2017).

These findings with the small sample of U.S. middle school teachers are similar to those of previous studies indicating that elementary teachers were not accurate at identifying their students' specific areas of difficulty (Bailey & Drummond, 2006; Hamilton & Shinn, 2003; Paleczek et al., 2017). It may be more efficient and less expensive to ask teachers to rate their students than to administer multiple reading tests (Kettler & Albers, 2013; Speece et al., 2010, 2011), but teacher judgments in the present study did not emerge as a viable alternative to collecting more detailed assessment data about students' reading abilities.

This is not unlike concerns expressed in studies of special education identification processes—that teachers were variable in their response to students and maintained self-defined, equivocal tolerance levels for students' performance (Gerber, 2005). Moreover, teacher judgments sometimes are used in research for either identifying participants or confirming that potential participants might benefit from a reading intervention being tested. Findings of the present study suggest researchers should be cautious about using teacher nomination as an eligibility criterion in the middle grades.

It is possible that informing teachers about the skills and the ways they are tested will improve the concordance between their ratings and objective measures (Bailey & Drummond, 2006; Südkamp et al., 2012). If this turns out to be the case, including teachers' judgments may become a reasonable way to address the data challenges faced by U.S. middle schools trying to address diverse student needs (O'Reilly et al., 2012). Alternatively, it might be more efficient to develop assessments that are capable of providing more specific information for teachers and negate the need for guesswork or specialized knowledge. To reduce the testing burden on teachers and students, there also is a need for measures that are automated, capable of pinpointing discrete skills for instructional starting points, and designed to be appropriate for middle school students.

Regardless of whether or not teacher judgments play a role in the process, identifying students is only a first step in planning and delivering differentiated instruction or targeted interventions. Teachers still may need extensive professional development and ongoing coaching to improve their instruction and the subsequent outcomes for students who need additional support (Brownell et al., 2017; Gerber, 2005).

Conclusion

The present study addressed several methodological gaps in the teacher judgment literature. Specifically, multilevel modeling was applied to explore how a sample of U.S. teachers' judgments about middle school students' reading skills predicted standardized test scores. Unfortunately, the lack of pattern in results across measures and skills suggests that the few moderate predictors identified were anomalous hits in a sea of misses. In the absence of disconfirming evidence from additional research or effective efforts to improve teachers' concordance, collecting teacher judgment data may not be a reliable means of identifying middle school students at risk of not reading proficiently or planning targeted instruction to address their needs.

References

Akram, K., Erickson, F., & Meyer, R. (2013). *Issues in the estimation of student growth percentiles* [Conference session]. Annual meeting of The Association for Education Finance and Policy, New Orleans, LA.

Bailey, A. L., & Drummond, K. V. (2006). Who is at risk and why? Teachers' reasons for concern and their understanding and assessment of early literacy. *Educational Assessment*, *11*, 149–178. https://doi.org/10.1207/s15326977ea1103&4_2

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bates, C., & Nettelbeck, T. (2001). Primary school teachers' judgments of reading achievement. *Educational Psychology, 21*, 177–187. https://doi.org/10/1080/01443410020043878

Begeny, J. C., Krouse, H. E., Brown, K. G., & Mann, C. M. (2011). Teacher judgments of students' reading abilities across a continuum of rating methods and achievement measures. *School Psychology Review*, *40*, 17. https://doi.org/10.1080/02796015.2011.12087726

Beswick, J. F., Willms, J. D., & Sloat, E. A. (2005). A comparative study of teacher ratings of emergent literacy skills and student performance on a standardized measure. *Education, 126*, 116–137.

Betjemann, R. S., Keenan, J. M., Olson, R. K., & DeFries, J. C. (2011). Choice of reading comprehension test influences the outcomes of genetic analyses. *Scientific Studies of Reading, 15*, 363–382. https://doi.org/10.1080/10888438.2010.493965

Brownell, M., Kiely, M. T., Haager, D., Boardman, A., Corbett, N., Algina, J., Dingle, M. P., & Urbach, J. (2017). Literacy learning cohorts: Content-focused approach to improving special education teachers' reading instruction. *Exceptional Children, 83*, 143–164. https://doi.org/10.1177/0014402916671517

Deeney, T. A., & Shim, M. K. (2016). Teachers' and students' views of reading fluency: Issues of consequential validity in adopting one-minute reading fluency assessments. *Assessment for Effective Intervention, 41*, 109–126. https://doi.org/10.1177/1534508415619905

Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test* (4th ed., Summary, Vol. 30). Pearson Education, Inc. https://doi.org/10.1037/t15144-000

Eason, S. H., Goldberg, L. F., Young, K. M., Geist, M. C., & Cutting, L. E. (2012). Reader-text interactions: How differential text and question types influence cognitive skills needed for reading comprehension. *Journal of Educational Psychology, 104*, 515–528. https://doi.org/10.1037/a0027182

Feinberg, A. B., & Shapiro, E. S. (2003). Accuracy of teacher judgments in predicting oral reading fluency. *School Psychology Quarterly*, *18*, 52–65. https://doi.org/10.1521/scpq.18.1.52.20876

Feinberg, A. B., & Shapiro, E. S. (2009). Teacher accuracy: An examination of teacher-based judgments of students' reading with differing achievement levels. *The Journal of Educational Research*, *102*, 453–462. https://doi.org/10.3200/JOER.102.6.453-462

Fien, H., Anderson, D., Nelson, N. J., Kennedy, P., Baker, S. K., & Stoolmiller, M. (2018). Examining the impact and school–level predictors of impact variability of an 8th grade reading intervention on at-risk students' reading achievement. *Learning Disabilities Research & Practice, 33*, 37–50. https://doi.org/10.1111/ldrp.12161

Firmender, J. M., Reis, S. M., & Sweeny, S. M. (2013). Reading comprehension and fluency levels ranges across diverse classrooms: The need for differentiated reading instruction and content. *Gifted Child Quarterly, 57*, 3–14. http://dx.doi.org/10.1177/0016986212460084

Foorman, B. R., Petscher, Y., Stanley, C., & Truckenmiller, A. (2017). Latent profiles of reading and language and their association with standardized reading outcomes in kindergarten through tenth grade. *Journal of Research on Educational Effectiveness, 10*, 619–645. https://doi.org/10.1080/19345747.2016.1237597

Francis, D. J., Barth, A., Cirino, P., Reed, D. K., & Fletcher, J. M. (2010). *Texas Middle School Fluency Assessment* (Version 2.0). University of Houston/Texas Education Agency.

Fuchs, L. S., Fuchs, D., & Compton, D. L. (2010). Rethinking response to intervention at middle and high school. *School Psychology Review, 39*, 22–28. https://doi.org/10.1080/02796015.2010.12087787

Geraci, M. (2014). Linear quantile mixed models: The lqmm package for Laplace quantile regression. *Journal of Statistical Software, 57*(13), 1–29. doi:10.18637/jss.v057.i13

Geraci, M. (2018). *lqmm: Linear quantile mixed models. R package, version 1.5.4.* https://cran.r-project.org/web/packages/lqmm/index.html

Gerber, M. M. (2005). Teachers are still the test: Limitations of response to instruction strategies for identifying children with learning disabilities. *Journal of Learning Disabilities, 38*, 516–524. https://doi.org/10.1177/00222194050380060701

Hamilton, C., & Shinn, M. R. (2003). Characteristics of word callers: An investigation of the accuracy of teachers' judgments of reading comprehension and oral reading skills. *School Psychology Review, 32*, 228–240. doi:10.1080/02796015.2003.12086195

Heller, R., & Greenleaf, C. L. (2007). *Literacy instruction in the content areas: Getting to the core of middle and high school improvement*. Alliance for Excellent Education.

Jaeger, E. L., & Pearson, P. D. (2017). The integration of Common Core and response to intervention: Supporting vulnerable readers in a time of sophisticated standards. *Educational Forum, 81*, 92–107. https://doi.org/10.1080/00131725.2016.1242676

Kettler, R. J., & Albers, C. A. (2013). Predictive validity of curriculum-based measurement and teacher ratings of academic achievement. *Journal of School Psychology, 51*, 499–515. https://doi.org/10.1016/j.jsp.2013.02.004

Koenker, R., & Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives, 15*, 143–156. doi:10.1257/jep.15.4.143

Kosanovich, M. L., Reed, D. K., & Miller, D. H. (2010). *Bringing literacy strategies into content instruction: Professional learning for secondary-level teachers*. RMC Research Corporation, Center on Instruction.

Lockwood, J. R., & Castellano, K. E. (2017). Estimating true student growth percentile distributions using latent regression multidimensional IRT models. *Educational and Psychological Measurement*, *77*, 917–944. https://doi.org/10.1177/0013164416659686

MacGinitie, W. H., MacGinitie, R. K., Maria, K., & Dreyer, L. (2000). *Gates-MacGinitie Reading Tests* (4th ed.). Riverside Publishing.

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. https://doi.org/10.1080/10349120500071886

Martin, S. D., & Shapiro, E. S. (2011). Examining the accuracy of teachers' judgments of DIBELS performance. *Psychology in the Schools*, *48*, 343–356. https://doi.org/10.1002/pits.20558

McKevett, N. M., & Kiss, A. J. (2019). The influence of data on teachers' judgments of students' early reading and math skills. *Psychology in the Schools*, *56*, 1157–1172. https://doi.org/10.1002/pits.22256

Missall, K. N., Hosp, M. K., & Hosp, J. L. (2019). Reading proficiency in elementary: Considering statewide testing, teacher ratings and rankings, and reading curriculum-based measurement. *School Psychology Review*, *48*, 267–275. https://doi.org/10.17105/SPR-2017-0152.V48-3

Nelson, P. M., Van Norman, E. R., & Lackner, S. K. (2016). A comparison of methods to screen middle school students for reading and math difficulties. *School Psychology Review, 45*, 327–342. https://doi.org/10.17105/SPR45-3.327-342

O'Reilly, T., Sabatini, J., Bruce, K., Pillarisetti, S., & McCormick, C. (2012). Middle school reading assessment: Measuring what matters under a RTI framework. *Reading Psychology, 33*, 162–189. https://doi.org/10.1080/02702711.2012.631865

Oslund, E. L., Clemens, N. H., Simmons, D., & Simmons, L. E. (2018). The direct and indirect effects of word reading and vocabulary on adolescents' reading comprehension: Comparing struggling and adequate comprehenders. *Reading and Writing: An Interdisciplinary Journal, 31*, 355–379. https://doi.org/10.1007/s11145-017-9788-3

Paleczek, L., Seifert, S., & Gasteiger-Klicpera, B. (2017). Influences on teachers' judgment accuracy of reading abilities on second and third grade students: A multilevel analysis. *Psychology in the Schools*, *54*, 228–245. https://doi.org/10.1002/pits.21993

R Core Team. (2017). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed., Vol. 1). Sage.

Reed, D. K. (2015). Middle level teachers' perceptions of interim reading assessments: An exploratory study of data-based decision making. *Research in Middle Level Education, 38*(6), 1-13. Retrieved from https://www.amle.org/portals/0/pdf/rmle/rmle_vol38_no6.pdf

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. https://doi.org/10.1177/0731948711432509

Speece, D. L., Ritchey, K. D., Silverman, R., Schatschneider, C., Walker, C. Y., & Andrusik, K. N. (2010). Identifying children in middle childhood who are at risk for reading problems. *School Psychology Review*, *39*, 258–276. doi:10.1080/02796015.2010.12087777

Speece, D. L., Schatschneider, C., Silverman, R., Case, L. P., Cooper, D. H., & Jacobs, D. M. (2011). Identification of reading problems in first grade within a response-to-intervention framework. *The Elementary School Journal, 111*, 585–607. doi:10.1086/659032

Stevenson, N. A. (2017). Comparing curriculum-based measures and extant datasets for universal screening in middle school reading. *Assessment for Effective Intervention, 42*, 195–208. https://doi.org/10.1177/1534508417690399

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, *10*, 743–762. https://doi.org/10.1037/a0027627

Valdez, A. (2013). Teacher judgment of reading achievement: Cross-sectional and longitudinal perspective. *Journal of Education and Learning*, *2*, 186–200. https://doi.org/10.5539/jel.v2n4p186

Woodcock, R. W., McGrew, K., & Mather, N. (2001). *Woodcock-Johnson Tests of Achievement* (3rd ed.). Riverside.