Hammill Institute on Disabilities

Focus on Autism and Other Developmental Disabilities
1–11
© Hammill Institute on Disabilities 2018
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1088357618762749
focus.sagepub.com
SAGE

Alternate Assessment Formats for Progress Monitoring Students With Intellectual Disabilities and Below Average IQ: An Exploratory Study

Francesca G. Jones, PhD[1], Diane Gifford, PhD[1], Paul Yovanoff, PhD[1], Stephanie Al Otaiba, PhD[1], Dawn Levy, MEd[1], and Jill Allor, EdD[1]

Abstract

As part of standards-based reforms, there is increasing emphasis on ensuring that students with moderate intellectual disabilities (ID), including students with Autism Spectrum Disorders (ASD), learn to read. There is also converging evidence that explicit teaching of letter sounds, phonics, and sight words is effective for this population, but that students' responsiveness varies. A critical part of individualizing reading instruction for students with disabilities is the reliable assessment of progress and mastery of reading skills. However, assessment of many students with ID and students with ASD is challenging because of attention, behavioral, and communication issues related to testing situation; therefore, obtaining consistent results often proves to be a difficult task. We hypothesized that alternate assessment presentation formats, as a testing accommodation, would improve the reliability, validity, and consistency of assessment performance. In this study, three different presentation formats—word lists, flash cards, and PowerPoint presentation—were used when administering proximal, curriculum-based reading assessments to determine whether a particular format increased student engagement, reduced the need for prompts, and increased accuracy of identifying known items on the test. While statistical analyses did not support the hypothesis of a format by student effect, visual analysis of the data did suggest that the number of prompts required varied by student as a function of assessment format. Most noteworthy, assessment reliability, estimated with generalizability theory, indicated that reliability increased as a function of format by student.

Keywords

assessment, autism spectrum disorders, literacy, mental retardation

It is troubling that most students with intellectual disability (ID) and students with Autism Spectrum Disorders (ASD) exit school with very limited reading ability (e.g., Wei, Blackorby, & Schiller, 2011), which limits opportunities for employment and general quality of life. However, recent research has shown that students with ID and ASD benefit not only from training on isolated skills such as sight word reading or letter-sound correspondences (e.g., Browder, Wakeman, Spooner, Ahlgrim-Delzell, & Algozzinexya, 2006 but also from intensive, individualized, comprehensive research-based reading instruction (Allor, Mathes, Roberts, Cheatham, & Al Otaiba, 2014; Browder, Ahlgrim-Delzell, Flowers, & Baker, 2012; Lemons & Fuchs, 2010).

While this research is promising, converging findings indicate large individual differences in response to intervention, with overall very slow growth on distal curriculum-based measures (CBM). For example, Allor et al. (2014) reported that even among students who received an intensive reading intervention, students in the mild ID range (IQs ranging 56–69) required about three academic years to progress from 10 words per minute (wpm) to 60 wpm on oral reading fluency. Students with IQs in the moderate range (40–55) needed longer time, approximately three and a half years, to move from 0 to 20 wpm.

Some recent studies have used CBM as well as proximal curriculum-based assessment of taught skills to assess reading growth for students with cognitive disabilities (Allor, Gifford, Al Otaiba, Miller, & Cheatham, 2013; Allor et al., 2014; Lemons, Mrachko, Kostewicz, & Paterra, 2012; Wallace, Tichá, & Gustafson, 2010). However, many assessment data points are needed to show a reliable trend

[1] Southern Methodist University, Dallas, TX, USA

Corresponding Author:
Francesca G. Jones, Department of Teaching and Learning, Simmons School of Education and Human Development, Southern Methodist University, Post Office Box 750455, Dallas, TX 75275-0455, USA.
Email: fjones@smu.edu

in growth because often there is great variance in performance even within one student's responses (Allor et al., 2013; Wallace et al., 2010). There are often compounding factors, such as students' behavior and attention span, which cause much variation in their performance (Ketterlin-Geller, 2008). This leaves the field with questions regarding how to effectively assess such variable growth on literacy skills (Lemons & Fuchs, 2010).

Need for Assessment Accommodations

Accurately assessing targeted skills such as letter sounds, word recognition, word attack, and word fluency is further complicated because students with ID and ASD may lack what Niebling and Elliott (2005) term *access skills*, those skills students need to show what they know about the target skills. For example, access skills for a timed progress monitoring assessment requires students to (a) understand a testing situation (sometimes with a less familiar examiner than their teacher), (b) attend to visual and auditory stimuli that may be different than familiar instructional materials, and (c) remain engaged throughout a timed task. In the colloquial, it can be difficult to discern "what they can do from what they won't do" and, as a result, finding accurate and sensitive assessments to progress monitor or measure mastery of literacy skills is a challenge (Baker, Spooner, Ahlgrim-Delzell, Flowers, & Browder, 2010; Wallace et al., 2010). In addition, students may have behavioral, language, or sensory challenges that also affect response and therefore the technical adequacy of assessment (Ketterlin-Geller, 2008).

Individuals With Disabilities Education Act (IDEA) and No Child Left Behind (NCLB) require practitioners to provide reliable and valid testing accommodations and modifications. While creating accommodations for assessments has received considerable attention from researchers for this population of students, the primary focus of any requirements for Universal Design of Assessment (UDA) has been on large scale accountability (Ketterlin-Geller, 2008; Ketterlin-Geller, Alonzo, Braun-Monegan, & Tindal, 2007). When the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education developed the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), the authors emphasized that when proper accommodations are not made for the assessment, the assessment is no longer fair or valid.

The National Center on Educational Outcomes defines accommodations for students with disabilities as, ". . . changes in testing materials or procedures that enable students to participate in assessments in a way that assesses abilities rather than disabilities" (National Center for Employee Ownership, 2016). Common accommodations are setting, timing, scheduling, presentation of stimuli, or required responses to ensure that students can show what they know about target skills by reducing issues related to access that are associated with a disability (Niebling & Elliott, 2005). Inherent in this definition of accommodation is that there is a differential effect for a student with a disability, but not a student without a disability. Unlike modifications, which change the content of the assessment, accommodations do not change what is being measured, only how it is being measured.

Accommodations for assessment are much more common for summative than for formative assessment. Interpretations of the results of a test that does not have proper accommodations may lead to invalid inferences and poor instructional decision making, which could diminish the effects of practitioners' ability to implement data-based individualization (DBI) of reading instruction (Danielson, Wexler, & Rosenquist, 2014; Lemons, Kearns, & Davidson, 2014). Along with educators, researchers also need reliable and valid measures to assess proximal skills targeted during intervention to develop and iteratively test interventions and to monitor participating students' responsiveness to determine the efficacy of a developed intervention.

The Current Study

This exploratory study was a part of the Institute of Education Sciences (IES) reading intervention development grant, Project Intensity. The aim of the grant was to examine the promise and feasibility of a series of books and activities specifically written to provide reading instruction and practice for children with ID and with borderline IQs, including students with ASD. To determine student progress in the program, we created proximal assessments of high frequency sight words. In format, the assessments were designed like word-level CBM; they were paper-based and words were presented in an array on a single page, but unlike many CBM, they were not timed. In the larger project, project staff administered these assessments weekly. Students were asked to read the list of words, untimed, and their score was the number of words read accurately.

The primary purpose of this exploratory study was to explore accommodations for progress monitoring that relate to the format of delivery; specifically, we compared our paper and pencil format to the same stimuli presented on flash cards or on PowerPoint slide presentation formats. We considered that different forms of the assessment might be more engaging, without changing the content. We had no hypotheses that one particular format would be most engaging for every student, but hypothesized that the preferred format would likely vary across students.

Table 1. Student Demographics.

Students	Gender	Ethnicity	Age	Diagnoses	IQ
Jacob[a]	Male	Caucasian	9	Down syndrome	47
Stephen[a]	Male	Caucasian	10	Down syndrome	53
Susie[a]	Female	Hispanic	11	Down syndrome	46
Elliot[b]	Male	Hispanic	8	Autism	71
Greg[b]	Male	Caucasian	8	Autism	78
Milton[b]	Male	Caucasian	5	Autism	70

[a]Attends School 1. [b]Attends School 2.

We were motivated to conduct this exploratory work following initial design trials because we found, as is typical for students with ID and with borderline IQs, that performance inconsistencies existed and that scores fluctuated widely across students and from day-to-day. One student in particular, Jacob, had widely variable performance and required extensive prompting to stay on task and attempt assessment items. It is important to note this study was conducted after the conclusion of the single case design study for the larger project. During this exploratory study, none of these students were receiving intervention on the target sight words; nor did assessors correct students when they made errors during testing. Consequently, there was no opportunity for the students to learn from the assessment over the course of the testing time period.

We addressed two research questions:

Research Question 1: What is the effect of a particular testing format on the number of items a student attempts, the number of required prompts to keep a student focused on the assessment, and on the percent of items correct?

Research Question 2: Is there an effect of format on measurement reliability?

Method

Setting

Two schools in a large metropolitan area of the Southwest participated in this study. The same schools were also part of the larger IES grant-funded project. School 1 was a private self-contained special education school for students with ID and ASD. School 2 was a public school that offered support to students with ID and ASD in various classroom settings, including a resource room, a self-contained classroom for students with autism, and a self-contained classroom for students needing intensive and functional learning experiences. In both schools, all classrooms had a teacher and one or more paraprofessionals who assisted in providing instruction and classroom management.

Participants

The participants included six students, ranging in age from 5 to 11 years. Three of the students attended School 1 and had diagnoses of Down syndrome. These students had IQ scores in the moderate range, between 40 and 54. The other three students attended School 2. Two were in general education classrooms and received special education services in a resource room. The other student received instruction in a self-contained classroom for students with autism. All three students had borderline IQs, between 70 and 79. Table 1 provides the student demographics for this exploratory study.

Measure

A proximal curriculum based assessment was developed by our research team to measure students' ability to read targeted sight words. This assessment consisted of 20 sight words. Fifteen of the words were randomly chosen from a list of words that had been specifically taught in the first portion of the curriculum. Five other untaught sight words were included from the Fuch's Word Identification Fluency measure, which includes words from the first grade Dolch word list (Fuchs, Fuchs, & Compton, 2004). Once the words were identified, word order was randomized to create three probes, each with a different word order.

For this study, we created accommodation through two alternate formats to the original paper and pencil format (PP): flash cards (FC) and PowerPoint slides (PPT). Across forms, words were kept in the same presentation order and written in the same font style (century gothic). On the paper-based form, the words were presented on one page in two columns of 10 words using a 42-point font size. For the other two forms, single words written in 150-point font size were centered on each of the flash cards and PowerPoint slides. In addition, the PowerPoint form was computer programmed to sound an audio ring tone when each new word was presented. Furthermore, each of the target words was shown for 5 s. Between these assessment items, blank slides with 2-s intervals were included. These three formats represented the testing conditions. Three alternate forms of every

Table 2. Order and Day of Testing by Groups of Students.

Order	Day 1	Day 2	Day 3
A	PPT	PP	FC
B	PP	FC	PPT
C	FC	PPT	PP

Note. PPT = PowerPoint slide; PP = paper-based; FC = flash card.

assessment format were also constructed. This design resulted in nine tests (three forms for each of three administration formats).

Design

To control for order effects, a matrix was created following a reduced Latin Square design (McKay & Wanless, 2005). This counter balance matrix has an identical order in the first row (n = 1, 2,. . .k) and the first column (n = 1, 2,. . .k). Within this reduced design, all rows and columns fall in the assigned natural sequence. This matrix decreases the chance of treatment order effect by removing it as a source of variability. To complete the matrix, we assigned format to day for the first row (n = PPT, PP, and FC) and then matched the sequence of format to order in the first column (see Table 2). To assign students to their starting order A, B, or C, we used a stratified sampling procedure and ranked them on reading ability to divide them into two groups of three. Then, within the two groups we alphabetized children by name and assigned them to a presentation order A, B, or C. As a result of this reduced Latin square design, students received each testing format (i.e., PPT = PowerPoint slide, PP = paper-based, and FC = flash card) once in any given order and only once on any given day across weeks.

Dependent Variables

The choice of dependent variables was based on the hypothesized effects of the assessment formats in the context of intervention research with the ID population. We hypothesized that accommodation using appropriate assessment format would (a) increase measurement reliability, (b) increase the number of items attempted, (c) decrease the number of prompts of redirections, and (d) increase overall performance. Collectively, the use of an appropriate accommodation in the form of administration format will increase the measurement technical adequacy in terms of these measurable qualities.

The primary dependent variable was the measure of taught words. A correct response was recorded and tallied using a frequency count when students accurately read the sight word presented on any of the three presentation forms. Only one correct response was recorded for each assessment item. Students could score up to 20 correct responses on any assessment form. Item self-corrections counted as correct responses. We calculated the percent correct as the number correct divided by the number attempted.

Furthermore, we examined whether there was an increase in items attempted. Items were considered as attempted when a student made eye contact with the probe and verbalized a response. In contrast, an item was scored as "not attempted" when students did not make eye contact with the probe, put their head on the desk, turned their face away, made no verbalization in response to assessment item, or otherwise refused to participate.

Finally, we measured the number of prompts or redirections. A prompt was coded as any redirection behavior exhibited by the assessor that was necessary to regain the student's attention or to get the child back on task. For this study, redirections included verbal prompts (e.g., "What word?") and nonverbal prompts (e.g., arm touching and facial expressions) to regain the student's attention. Using videotapes to observe assessors' behaviors, coders tallied redirections using a frequency count for each of the testing periods.

Procedures

On each testing occasion, trained assessors took individual students to a quiet space with minimal distractions at their respective schools. Students were then presented with the assessment in the format order as predetermined by the reduced Latin square design. Students were asked to read each word to the best of their ability. All assessment sessions were videotaped using iPads.

At a later time, the trained team members watched the videos to code items attempted. They also coded the number of prompts or redirections per session used by the assessors to keep a student on task during the testing occasion. Finally, the percentage correct (number correct/number attempted) was calculated.

Interobserver Reliability

Assessor reliability.

Prior to the administration of the proximal measures to the students, two research team members, one with a doctoral degree and one with a master's degree, both in Education, participated in assessment training. These team members then collected student data over the course of the 3-week period. Both were trained in the administration of the assessment and practiced to obtain reliable delivery and scoring. During training, they achieved an average reliability score of 93%.

Video coding reliability.

Three research assistants, one with a doctoral degree and two with master's degrees in Education, were trained and coded the 54 videos (i.e., nine testing sessions for each of the six participants). Coder reliability was calculated on three occasions to ensure the alignment of student behavior identification. Overall, an average of 93% reliability was achieved.

Data Analysis

Analyses were completed to test hypotheses of an administration format effect on (a) number of items attempted, (b) number of prompts necessary to keep the student on task, and (c) percent of items which were responded to correctly. In addition to these variables, we analyzed the impact of format on measurement reliability. Recall that each of the students was tested with each of the three formats on three independent occasions controlling for order. Treating repeated measures of each student as a blocking factor within each format, and testing for format effect, a nonparametric analysis was completed to test for a main effect of format, a main effect of student, and an interaction effect on each of the three dependent variables. Nonparametric statistical tests were used because of the small number of cases and the widely variable and nonnormal distribution of assessment scores. After ranking all scores, using procedures described by Shirley (1987), the mean rank for the format and student was compared statistically using a Kruskal–Wallis test of the mean ranks. In addition to the main effects of format and student, the interaction effect was used to test the hypothesis that format effects are conditional on student.

Additional analyses were completed to test the hypothesis of administration format on measurement reliability. Using generalizability theory (Brennan, 2000; Yovanoff, Tindal, & Geller, 2010), reliability was estimated by computing the systematic variation due to students, format, and occasions, and the interaction of these sources of variability. While student true score variability is desirable, other sources of variability are regarded as systematic and random “error” (attributable to sources such as occasions, format, student). generalizability theory focuses closely on error variance, and extending beyond classical test theory, provides an estimate of observed score variability attributable to administration format.

Results

Our exploratory study was designed to explore whether there was an effect of format on items attempted, the number of prompts required, and the percent correct on a proximal CBA. We also explored the effect of format on measurement reliability.

Effects on Items Attempted, Required Prompts, and Percent Correct

To address the first research question regarding the effect of a particular testing format on the number of items a student attempted, the number of prompts required, and the percent of items correct, we analyzed data for the six students who were measured on nine occasions. Figures 1, 2, and 3 present graphically the scores for (a) number of items attempted, (b) number of required prompts, and (c) percent correct adjusted for number of items attempted, respectively. Within format, the open circles are each of the three independent measures and the solid circle is the mean score. Table 3 provides a summary of the analysis of the Kruskal–Wallis ranks test of factorial effects based on procedures described by Shirley (1987). The chi-square test statistic is used to evaluate the format, student, and format-by-student effects. For each dependent variable, the hypothesized format-by-student interaction effect was not obtained. Only student’s effects were significant statistically. These effects are apparent in the graphs.

Effects on Measurement Reliability

Our second research question pertained to measurement reliability and was based on the observation that measures of reading (observed percent correct) for any individual appear to vary differentially depending on format. Figure 3 provides graphs of the observed percent correct. The within-student variability across format across student is striking. Using generalizability theory procedures, we have focused on two relevant issues, (a) measurement reliability and (b) sources of variance. We used a person by occasion nested in format ($p \times o{:}f$) design to estimate measurement reliability, treating format as fixed. (Occasion refers to the repeated measurement of each student.) The estimated generalizability coefficient (reliability) is 0.86.

Though measurement reliability is very high, we looked closely at the variance components on which reliability is based. One hypothesis that a significant source of error variance is occasion-by-format conditional on person ($o{:}f$ conditional on p). The repeated measures should remain stable. We anticipate, however, that this will depend on the measurement format. And, this conditional variation will depend on the examinee if measurement format does function as an accommodation. Table 4 provides a summary of the relevant variance components. While 54.85% is attributable to persons (true score variance), a nontrivial 26.5% is due to the variation of measures within a format conditional on students.

Figure 3 provides a graphic illustration of the variability of proportion correct scores across occasion (three repeated measures) within a format for each student. Ideally, for each student we expected small variation across his or her nine measures (three formats each with three measures), which translate into high reliability. It is apparent visually, however, that for each student the variability depends on format. For instance, Elliot’s scores vary relatively less when using flash cards (FC). For Susie, we see relatively large variation (low reliability) when using paper and pencil (PP) administration. Based on the generalizability theory results reported in Table 3 and visual analyses of Figure 3, it is apparent that measurement reliability does depend upon administration format and student.

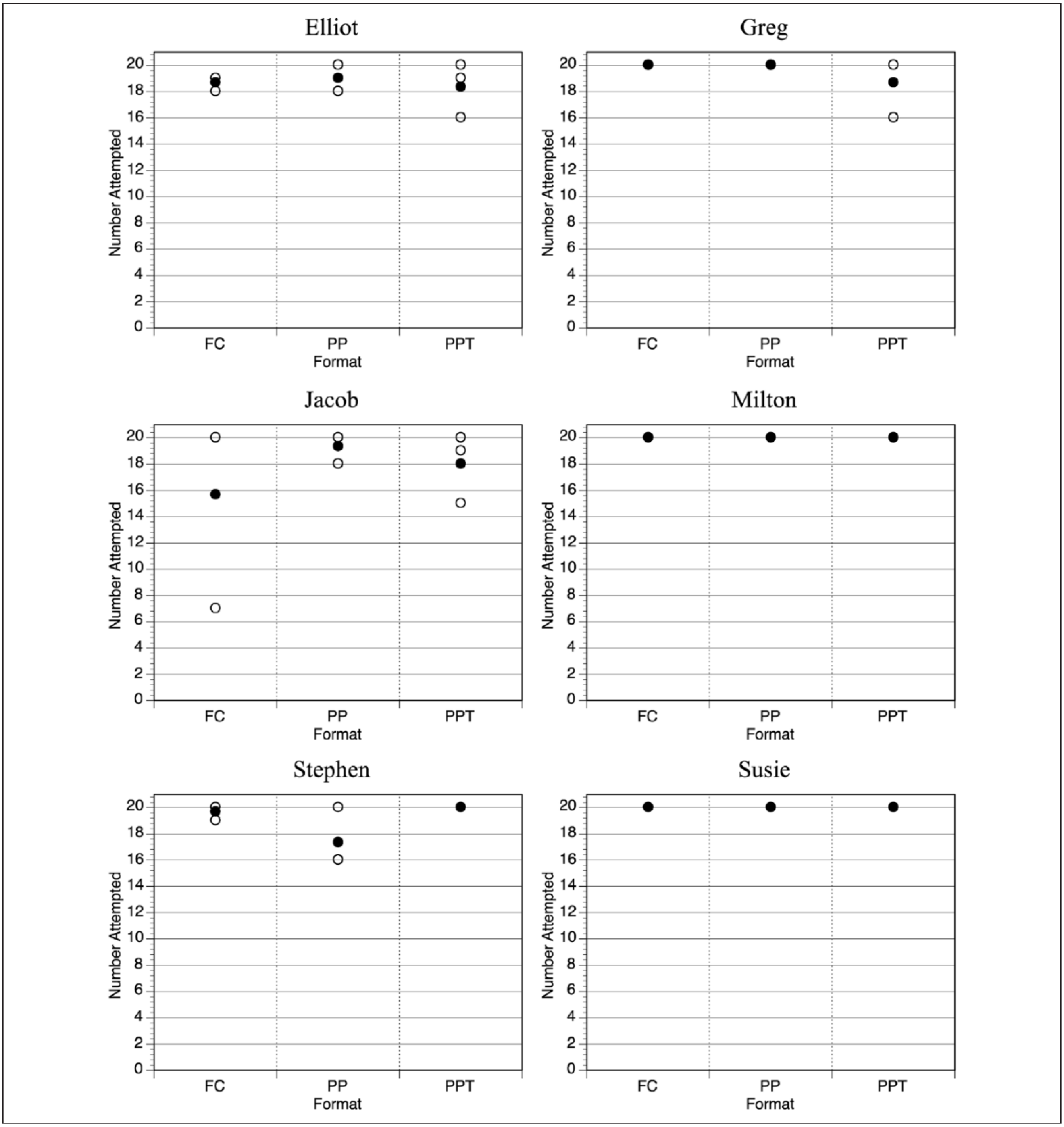

Figure 1. Number of items attempted by format by case (○ is one of three measures per format, ● is the mean of those three measures).

Discussion

We designed this exploratory study to extend the limited research based on progress monitoring for students with disabilities by examining accommodations for progress monitoring measures, in this case, a proximal CBA of taught sight words. We hypothesized that different formats of the assessment would affect the number of items a student attempted, the prompts required to keep them on task, and the percent of items they respond to correctly. In addition, we examined the effect of format on measurement reliability. This discussion begins with a summary of results related to the research questions; next we discuss in greater depth the variable performance of Jacob, whose variable

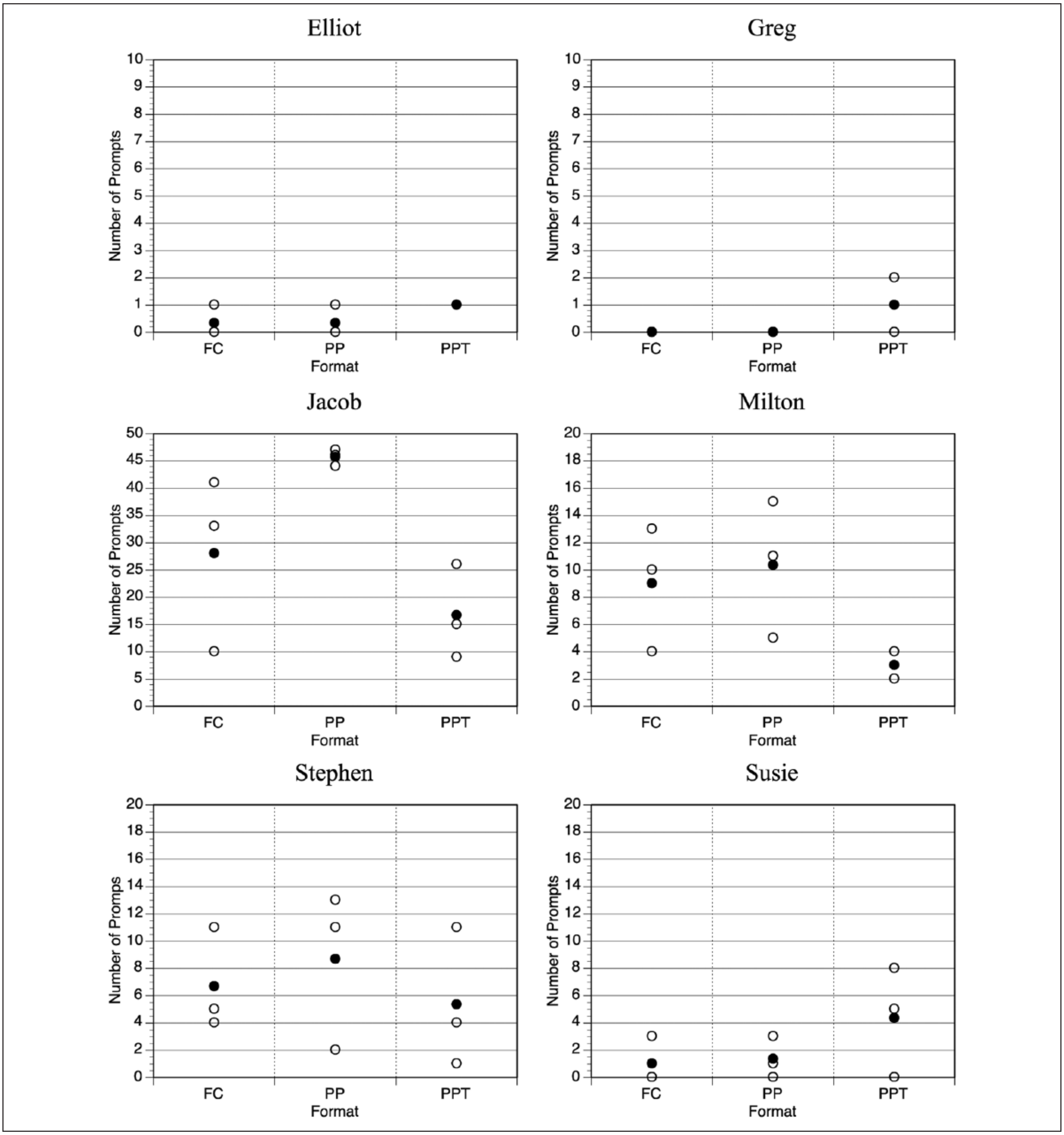

Figure 2. Number of prompts by format by case (○ is one of three measures per format, ● is the mean of those three measures).

assessment performance inspired the study. Then we discuss preliminary study implications, limitations, and directions for future research.

Our first research question addressed the effect of a particular testing format on the number of items a student attempts, the number of required prompts to keep a student focused on the assessment, and on the percent of items correct. Our analyses indicated that, on average, there was no clear pattern of one format emerging as optimal for any of these three outcomes. These findings should be considered preliminary given the design of the study and the small number of participants. However, our results add uniquely to the research because it is the first study to explore accommodations for CBA in the context of progress monitoring

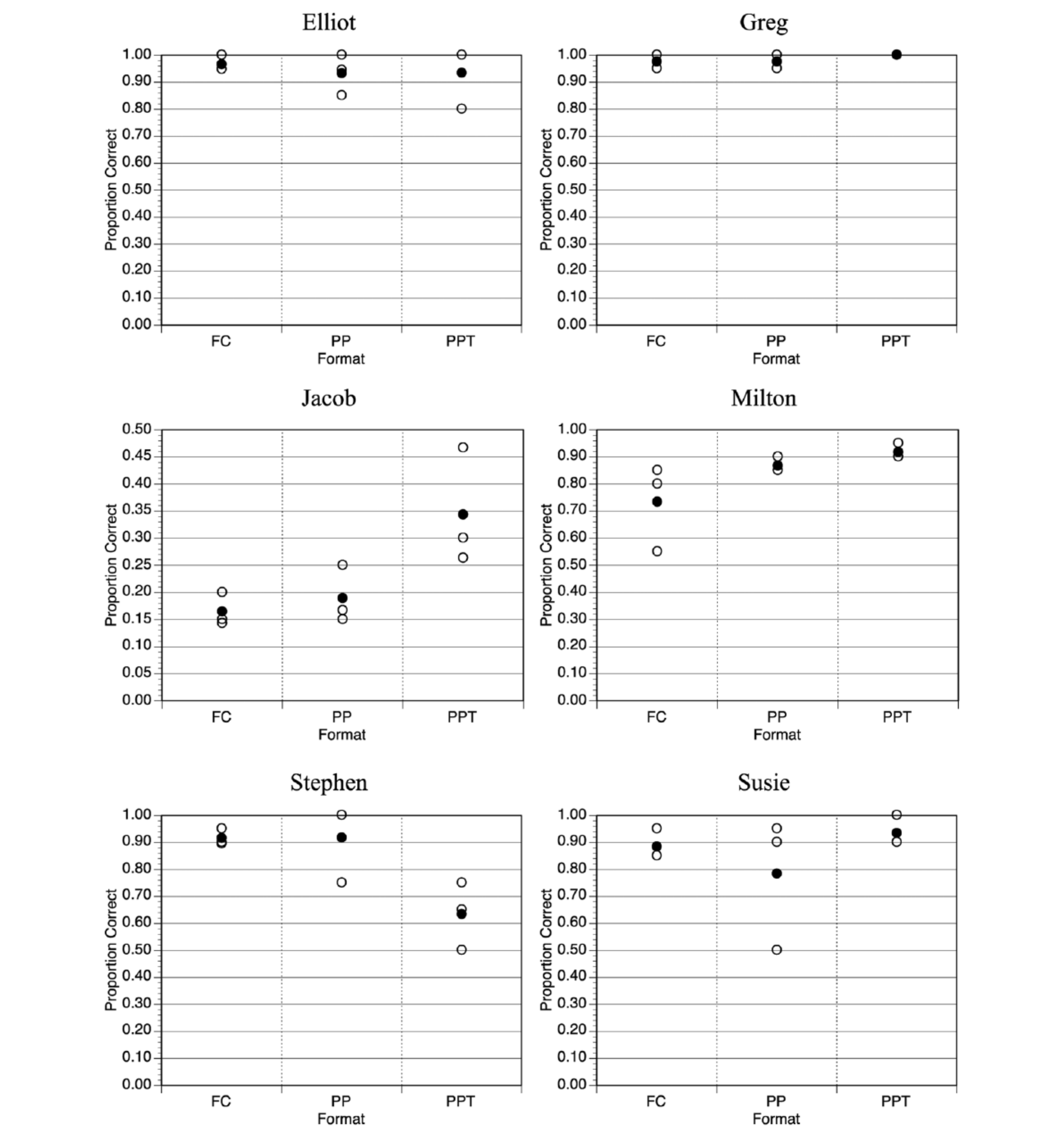

Figure 3. Proportion correct by format by case (○ is one of three measures per format, ● is the mean of those three measures).

for students with ID and low IQs. In addition, it is the first study to examine not only percent correct but also items attempted and the number of prompts required.

As shown in Figure 1, which shows the number of assessment items attempted by students, some students attempted the same number of items across all formats, while others demonstrated markedly different performance across formats. In contrast, in Figure 2, there was slightly greater variation in the number of prompts required by format. Three students, Elliot, Greg, and Susie, showed relatively low levels of prompting needed across formats. However, the other three students, Jacob, Milton, and Stephen, who required a higher number of prompting overall, also showed more variability across and within format.

Table 3. Number of Items, Number of Prompts, and Percent Correct Format by Person Kruskal–Wallis Ranks Test Summary.

Source	Sum of Squares	*df*	χ^2	*p*
Number of items				
Corrected model	3,345.92	17	21.09	.22
Format	24.08	2	0.15	.93
Student	2,243.20	5	14.13	.02
Format by student	832.12	10	5.24	.87
Corrected total	8,094.00	51		
Number of prompts				
Corrected model	8,394.51	17	38.91	.002
Format	105.52	2	0.49	.78
Student	7,743.39	5	35.00	0.00
Format by student	697.94	10	3.24	.98
Corrected total	10,787.50	50		
Percent correct				
Corrected model	8,758.67	17	38.79	.002
Format	47.43	2	0.21	.90
Student	7,714.93	5	33.17	.001
Format by student	996.60	10	4.41	.93
Corrected total	11,516.50	51		

Table 4. Variance Component Estimates for Person × Observations Nested in Format.

Source	*n*	Estimated variance component	*SE*	95% confidence interval	% total variability
p	5	0.063	0.035		54.85
f	2				
o;f	6				
p × f	10	0.023	0.014		19.66
p × o;f	30	0.031	0.008		26.50

Figure 3, which displays the percent of items correct, shows marked variability across student and format. Elliot and Stephen demonstrated not only a greater number of items scored correct but also a more consistent performance overall for the flash card format of presentation. For Milton, his highest and most consistent performance was in the PowerPoint format. Greg displayed a consistently high and reliable performance across all formats. Susie had similar performance for the flash card and PowerPoint formats and markedly lower performance and more variability in the paper and pencil format. We address Jacob's pattern of performance in greater detail later.

Figure 3 also reflects the format effect on measurement reliability. The results of the statistical analysis demonstrated that, while a format was not individually predictive of a student's performance, format does appear to affect the consistency of a student's score, rendering more reliable measurement. Over half of the error variance was attributed to within-student differences, while over a quarter of the variance was due to the format within student. This finding adds uniquely to the literature on universal design for assessment related to the reliability of accommodations for progress monitoring for this population of students.

Moreover, our findings about the variable performance on our CBA converge with findings from other studies that have used CBM or CBA with this population (i.e., Allor et al., 2013; Lemons et al., 2012; Wallace et al., 2010). Notably, we also learned that there were marked variations within each individual within each format. These findings are encouraging because we were able to reduce the variance in performance by changing the administration format.

Furthermore, what we found is consistent with what Anderson, Farley, and Tindal (2015) found in their work with standardized assessments, that these accommodations provided our students access to the assessments and did not compromise, but rather improved, reliability. They had stated that one concern with accommodations is that they can be administered unreliably and increase the unreliability of the exam; this concern was particularly high for accommodations where the administrator could be a factor. However, given the accommodations we provided, a change in format only, we found that reliability of performance was actually increased.

Jacob: A Case of Variability

We designed this exploratory study due to concerns about the extreme variability in Jacob's scores and the high number of prompts he required to stay on task to attempt items on the assessment. It is noteworthy that his pattern of performance in this study is consistent with his performance in the larger study. For him, the number of items attempted was high and relatively consistent for the paper and PowerPoint formats, but very inconsistent and even low for the flash card format.

In Figure 2, one can see that Jacob required more prompts than any other student in the study that the number of prompts he needed did seem to be affected by format. The paper and pencil format led to the highest number of prompts (ranging from 43–47) during each testing occasion. The PowerPoint format required fewer prompts but led to more variance (range from 10–26) overall. The flash card format required a wide variation of prompts (range from 10–42) as well.

Figure 3 indicates that the format (PPT) that elicited Jacob's highest mean performance was also his most variable. In other words, Jacob had a higher percent correct with the PowerPoint format, but at the same time he demonstrated his most inconsistent performance (range, 26%–46%) across PPT testing occasions. In contrast, his most consistent scores, and his lowest scores, were those he earned using the flash card format (range, 15%–20%). These data are a challenge to interpret. These findings prompt the question, what then would be considered Jacob's "best" performance and which format should be used for his future progress monitoring? On one hand, from the point of view of teachers or school psychologists, the best performance might be interpreted as the format that allowed him to demonstrate the highest number of words correct; in his case the PowerPoint format (35%). On the other hand, from a measurement perspective, the best performance might be the most reliable, the flash cards (16%). Certainly, he knew the fewest words of all the six students, regardless of format and so his "best" is clearly relative.

Implications for Practice

There are several important, albeit tentative, implications of our study for teachers, school psychologists, and diagnosticians. The first implication is that it is possible to use UDA principles to create accommodations for formative assessments like CBA for this population of students. As no single format emerged as optimal for all students, professionals may need to explore a variety of formats to determine which works best for a student. The formats we used are simple and accessible to all teachers.

Our study affirms Ketterlin-Geller (2008) and the expressed need under IDEA for students with disabilities to have universal design for assessment through accommodations. In other words, students with low IQs must have an accommodation that ameliorates the impact on student performance. Providing student's access to CBA via format options is one way to increase access and student success. These accommodations need to be considered for not only high stakes, summative assessments, but also for formative CBM and CBA so that teachers have accurate data throughout the year on which to base instructional decisions.

A second implication relates to measurement reliability, namely, if a teacher can increase the reliability and validity of CBA assessment by changing the format of assessment administration, this could lead to more accurate use of the data to guide instructional decisions. It is important for practitioners to know that up to a quarter of the variance in a student's scores can be attributable to format. These results may not be surprising to seasoned practitioners who have wrestled with the "can't do vs. won't do" testing conundrum across their careers. If we can determine an accommodation that results in a more consistent performance, then we can determine the "best" accommodation. Even if this accommodation does not initially lead to the highest performance by the student, if it leads to the most consistent performance, then we know we have more accurately captured that students' true score and, therefore, make more reliable instructional decisions. This finding is important because of what we know about large individual differences in student performance of students with low IQs (Allor et al., 2013; Wallace et al., 2010).

Limitations and Directions for Future Research

As with most school-based research, there were challenges to the study, which limit our findings. First and foremost, though our design was experimental, this was an exploratory study involving relatively few testing occasions and a small number of participants. However, the findings suggest that future replication studies with larger samples and over a longer period of time are warranted. A second important limitation is that we measured only sight words taught; thus, additional work is needed to learn whether our findings generalize to other types of items, such as decodable words or letter names. In addition, future research should explore transfer to more distal CBM measures. A third limitation related to Jacob and the conundrum of not being able to easily determine which was the "best" format for a student with relatively low knowledge and relatively high off-task behavior. This points to the need to possibly consider more accommodations of format, access, or behavior for students to attain their most reliable and accurate performance.

Another direction for research relates to the design of assessments. Researchers and test publishers may consider the need for alternative formats as they design proximal measures for progress monitoring and intervention research. Right now, many assessments are available only through paper and pencil administration and some on the computer, which still requires student computer literacy and physical skills. Ideally, test producers may offer assessments in a variety of formats to simplify this process for practitioners.

Finally, teachers will likely need professional development and guidelines for selecting testing format accommodations for their students. In fact, special educators and school psychologists would likely benefit from learning more about UDA principles and the need for appropriate accommodations for this population. Furthermore, once practitioners are able to obtain more reliable data, they will need continued support and training to ensure that they know how to use those data to make instructional decisions for their students.

In conclusion, despite the study limitations, this exploratory study demonstrated the need for and promise of CBA testing accommodations. Findings indicate the need for future research to determine how to better assess students with low IQs to ensure that we are capturing their best and most reliable performance. We recognize that the field is just beginning to address the challenge of assessing students with low IQs; however, this study is a promising start to determining how to better assess students.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Support for carrying out this research was provided by grant R324A130102 from the Institute of Education Sciences. The views expressed herein are those of the authors and have not been reviewed or approved by the granting agency.

References

Allor, J. H., Al Otaiba, S., Yovanoff, P., Cheatham, J., Gifford, D., Levy, D., . . . & Jones, F. (2015, July). *The effects of a text-centered supplemental curriculum for students with intellectual disabilities*. Annual Meeting of the Society for the Scientific Study of Reading, Kona, HI.

Allor, J. H., Gifford, D. B., Al Otaiba, S., Miller, S. J., & Cheatham, J. P. (2013). Teaching students with intellectual disability to integrate reading skills: Effects of text and text-based lessons. *Remedial and Special Education*, *34*, 346–356. doi:10.1177/0741932513494020

Allor, J. H., Mathes, P. G., Roberts, J. K., Cheatham, J. P., & Al Otaiba, S. (2014). Is scientifically based reading instruction effective for students with below average IQ's? *Exceptional Children*, *80*, 287–306.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Anderson, D., Farley, D., & Tindal, G. (2015). Test design considerations for students with significant cognitive disabilities. *Journal of Special Education*, *49*, 3–15.

Baker, J. N., Spooner, F., Ahlgrim-Delzell, L., Flowers, C., & Browder, D. M. (2010). A measure of emergent literacy for students with severe developmental disabilities. *Psychology in the Schools*, *47*, 501–513.

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement*, *24*, 339–353.

Browder, D. M., Ahlgrim-Delzell, L., Flowers, C., & Baker, J. (2012). An evaluation of a multicomponent early literacy program for students with severe developmental disabilities. *Remedial and Special Education*, *33*, 237–246.

Browder, D. M., Wakeman, S. Y., Spooner, F., Ahlgrim-Delzell, L., & Algozzine, B. (2006). Research on reading instruction for individuals with significant cognitive disabilities. *Exceptional Children*, *72*, 392–408.

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

Fuchs, D., Fuchs, L. S., & Compton, D. (2004). Identifying reading disabilities by response to intervention: Specifying measures and criteria. *Learning Disabilities Quarterly*, *27*, 216–227.

Ketterlin-Geller, L. R. (2008). Testing students with special needs: A model for understanding the interaction between assessment and student characteristics in a universally designed environment. *Educational Measurement: Issues and Practices*, *27*, 3–16.

Ketterlin-Geller, L. R., Alonzo, J., Braun-Monegan, J., & Tindal, G. (2007). Recommendations for accommodations: Implications for (in)consistency. *Remedial and Special Education*, *28*, 194–206.

Lemons, C. J., & Fuchs, D. (2010). Modeling response to reading intervention in children with Down syndrome: An examination of predictors of differential growth. *Reading Research Quarterly*, *45*, 134–168.

Lemons, C. J., Kearns, D. M., & Davidson, K. A. (2014). Data-based individualization in reading. *Teaching Exceptional Children*, *46*(4), 20–29.

Lemons, C. J., Mrachko, A. A., Kostewicz, D. E., & Paterra, M. F. (2012). Effectiveness of decoding and phonological awareness interventions for children with Down syndrome. *Exceptional Children*, *79*, 67–90.

McKay, B., & Wanless, I. (2005). On the number of Latin squares. *Annals of Combinatorics*, *9*, 335–344.

National Center on Educational Outcomes. (2016, June). *Accommodations for students with disabilities*. Minneapolis: University of Minnesota. Retrieved from https://nceo.info/Resources/publications/TopicAreas/Accommodations/Accomtopic.htm

Niebling, B. C., & Elliott, S. N. (2005). Testing accommodations and inclusive assessment practices. *Assessment for Effective Intervention*, *31*, 1–6.

Shirley, E. A. C. (1987). Applications of ranking methods of multiple comparison procedures and factorial experiments. *Applied Statistics*, *36*, 205–213.

Tindal, G., Yovanoff, P., & Geller, J. P. (2010). Generalizability theory applied to reading assessments for students with significant cognitive disabilities. *Journal of Special Education*, *44*, 3–17.

Wallace, T., Tichá, R., & Gustafson, K. (2010). Technical characteristics of general outcome measures (GOMs) in reading for students with significant cognitive disabilities. *Reading & Writing Quarterly: Overcoming Learning Difficulties*, *26*, 333–360.

Wei, X., Blackorby, J., & Schiller, E. (2011). Growth in reading achievement of students with disabilities, ages 7 to 17. *Exceptional Children*, *78*, 89–106