Article

Hammill Institute on Disabilities

Assessment for Effective Intervention
2018, Vol. 43(2) 96–109

DOI: 10.1177/1534508417727489
aei.sagepub.com
SAGE

The Importance of Replication in Measurement Research: Using Curriculum-Based Measures With Postsecondary Students With Developmental Disabilities

John L. Hosp, PhD[1], Jeremy W. Ford, PhD[2], Sally M. Huddle, PhD[3], and Kiersten K. Hensley, PhD[4]

Abstract

Replication is a foundation of the development of a knowledge base in an evidence-based field such as education. This study includes two direct replications of Hosp, Hensley, Huddle, and Ford which found evidence of criterion-related validity of curriculum-based measurement (CBM) for reading and mathematics with postsecondary students with developmental disabilities (DD). Participants included two cohorts of postsecondary students with DD enrolled in a 2-year certificate program at a large Midwestern university (n = 24 and 21). Using the same standardized procedures as Hosp et al., participants were administered CBMs for Oral Passage Reading (OPR), Maze, Math Computation, and Math Concepts and Applications. Descriptive statistics and bivariate correlations between CBMs and the content-appropriate *Woodcock–Johnson Tests of Achievement–Third Edition* were calculated. No significant differences in criterion-related validity coefficients between cohorts were found but differences between the correlations for Math Computation and Math Concepts and Applications identified in Hosp et al. were not found in either replication cohort.

Keywords

curriculum-based measurement, developmental disabilities, postsecondary, replication

In 2002, two pieces of legislation put an increased emphasis on the establishment of an evidence base for the field of education. First, the reauthorization of the Elementary and Secondary Education Act (which renamed it as No Child Left Behind or NCLB) emphasized the use of "scientifically based research" as the basis for educational practices citing the term well over 100 times (No Child Left Behind Act, 2002). On February 6, 2002 Susan Neuman, then the assistant secretary for Elementary and Secondary Education at the U.S. Department of Education hosted a seminar on scientifically based research which framed the term and its use as the foundation for identification of effective practices in education (Neuman, 2002). Second was the signing of the Education Sciences Reform Act (ESRA) in November of that year which authorized the Institute for Education Sciences (IES) and established its mission to provide leadership in expanding fundamental knowledge of education which includes effectiveness of educational practices (Education Sciences Reform Act, 2002). At the time, IES defined randomized controlled trials as the gold standard in a presentation to the American Educational Research Association by Russ Whitehurst, the first director of IES (Whitehurst, 2003). Because of the shift in emphasis to efficacy of educational practices and programs, this was not an unexpected turn; however, the emphasis on generation of new knowledge has contributed to a "crisis of (non-) replication" (Cook, 2014, p. 233).

Replication is often considered to be at the heart of scientific inquiry (Schmidt, 2009), and education is a field which prizes inquiry. Moreover, replication is important because it is a means of controlling for threats to both internal and external validity that are inherent in any single research study by allowing for a systematic examination of

[1]University of Massachusetts Amherst, USA
[2]Boise State University, ID, USA
[3]Mesa County Valley School District 51, Grand Junction, CO, USA
[4]Minnesota State University, Mankato, USA

Corresponding Author:
John L. Hosp, PhD, Department of Student Development, University of Massachusetts Amherst, S164 Furcolo, Amherst, MA 01003, USA.
Email: johnhosp@umass.edu

the impact of those threats across studies (Schmidt, 2009). It is this aggregation that allows for the impartation of the term "evidence based" because through repeated study, the convergence of findings and understandings can be examined. This is a fundamental component of the scientific method (Gall, Gall, & Borg, 2006). Although the importance of replication is well-known, there remains little work detailing specifics of replication.

Although there are a few conceptual frameworks for types of replication (see Schmidt, 2009 for a review), a common current classification compares direct replication with conceptual replication. Direct replication uses the exact study design, sampling, and methods as the original research but can be conducted by the same or independent researchers. It is designed to examine the accuracy of specific results and data from the original study (Makel & Plucker, 2014). In contrast, conceptual replication uses different design, sampling, methods, and/or analysis to more fully explore the construct under study rather than the methods or data (Makel & Plucker, 2014). To date, when a study can be considered replicative, it generally is a conceptual replication so it can be framed as new (avoiding publication bias) and the results can be aggregated using meta-analytic methods. The different types of replication (i.e., direct and conceptual) are important to consider within the functions they serve.

Schmidt (2009) outlined five different functions of replication to (1) control for sampling error, (2) control for research artifacts, (3) control for fraud, (4) generalize findings to a different population, and (5) verify underlying hypotheses. Although the functions do not align perfectly with direct and conceptual replication, Functions 1 to 3 tend to focus on the specific results of the original study (similar to direct replication), whereas Functions 4 to 5 focus on extending the findings or construct from the original study. Both are important, but the latter are more commonly represented in the published literature. See Makel et al. (2016), Lemons et al. (2016), Cook, Collins, Cook, and Cook (2016), and Therrien, Mathews, Hirsch, and Solis (2016) for recent reviews of replication in special education.

A common theme of these four reviews and recent calls for increased replication in education (Cook, 2014; Makel & Plucker, 2014) is an exclusive focus on intervention research. This focus is likely for a few reasons. First is the emphasis on evidence-based practices and efficacy that are the current context of education in the United States (see above). Second, in order for a practice to be considered evidence based, at least four experimental or quasi-experimental study replications (Gersten et al., 2005) or five single case study replications (Horner et al., 2005) should be present as a condition for "evidence based." Third, a foundational principle of the U.S. education system is that teaching can effect learning and that learning is the heart of education (Webb & Metha, 2016). Similar to evidence-based practices in medicine which works on the belief that physicians using best practices increase patient health, in education, we generally operate with the belief that a teacher using effective practices will increase student learning. Within this context, this emphasizes that evidence-based intervention practices are crucial. However, replication is also important to consider in assessment research because if our measurement tools and procedures are not as rigorously evidence based as the interventions and protocols, it introduces a source of error into research studies that will undermine confidence in the foundation of our evidence base.

Schmidt's five functions of replication are just as crucial for assessment research as intervention research. Schmidt's first three functions are important for the same reasons in assessment research as intervention research—to control for threats to the validity of inferences from the findings of a single research study. Sampling error or research artifacts (such as a close alignment of the instruction/intervention students receive and the content of the assessment tools) can affect assessment research. Functions 4 and 5 are also important because it is important that the theory of the construct under measurement is not dependent upon the sample and that there is a consistency in measurement for different groups of students. In addition, classical test theory holds that any observed score from administration of an assessment instrument is composed of two components—the true score and measurement error (Crocker & Algina, 2006). Ways of combating measurement error and therefore increasing the reliability of measurement include adding items to a scale, having multiple scorers score the instrument, and increasing the frequency of administration through alternate forms, readministration of the same form, or ongoing progress monitoring. All of these are essentially methods of replicating a finding to increase the confidence in the veracity of the result. Research into the technical adequacy of an instrument can benefit equally from replication, and there are some examples of syntheses demonstrating this utility. These syntheses serve Functions 4 and 5 well as they can control for sources of error across studies to look for patterns of consistency or inconsistency; however, direct replications are still needed to address Functions 1 through 3.

Reschly, Busch, Betts, Deno, and Long (2009) conducted a meta-analysis of 41 studies that included at least one correlation between Oral Passage Reading (OPR) and a criterion-reading measure. None of the included studies are direct replications as all differed from each other on at least one aspect of design (e.g., sampling, participant characteristics, criterion measure) and in many cases more than one. These conceptual replications allow a fuller examination of the construct of OPR—particularly when analyzed meta-analytically so that mediating or moderating variables and publication bias could be examined empirically—but ignore the facets of method and design that may also impact the results.

In math, Foegen, Jiban, and Deno (2007) examined 32 studies of curriculum-based measurement (CBM). Although their synthesis did not include a meta-analysis, they examined the correlations between M-CBM and a variety of criterion measures in grades ranging from preschool to secondary grades. No two studies in their review are direct replications but rather were conceptual in nature varying on one or more study characteristics. Similar to the Reschly et al. (2009) review, this provides important information about the potential range of correlations and generalization to different populations; however, without direct replication of any of the studies it is difficult to make inferences about the robustness of the findings. What is needed to extend the literature on replications in CBM research is a direct replication using the same methods and materials but with a different sample of students. This will provide for an examination of the random error present in assessment while attempting to control for all other variables. Conceptually, this is similar to single case experimental designs (Kennedy, 2005) in that control is maintained over all variables other than the one being experimentally manipulated. This study includes such a comparison using CBMs to build off of the syntheses that have been conducted. This adds direct replication to the existing literature of more conceptual replications. Our focus was with postsecondary students with developmental disabilities (DD) because there is a lack of research on using CBM with this population; therefore, we next provide a rationale for their use.

Why Use CBM With Postsecondary Students With DD?

There is an increasing amount of postsecondary options for students with DD, with over 220 postsecondary programs for students with DD in the United States and Canada (Think College, 2014). Postsecondary education includes any education after the high school level, including vocational training, community college, and 4-year college settings. Programs for individuals with DD usually fall into one of three categories: hybrid model, substantially separate model, and inclusive individual support model. The hybrid model allows for students with DD to participate in both social events and academic courses with nondisabled peers, as well as coursework designed specifically for students with DD. In a substantially separate model, students with DD might participate in social activities with nondisabled peers, but courses are taken only with other students with disabilities. An inclusive individual support model is driven by the individual student's academic and career goals, and generally does not include a campus program designed specifically for students with DD (Hart, Grigal, Sax, Martinez, & Will, 2006).

Regardless of the model, a majority of the programs focus on an outcome of gaining quality employment (Papay & Bambara, 2012). Basic proficiency in reading and mathematics is related to better employment options. Quality jobs with the highest growth rates require reading and mathematical skills, leaving struggling readers with fewer options for gainful employment (Liming & Wolf, 2008; National Joint Committee on Learning Disabilities [NJCLD], 2008). Insufficient reading skill and high rates of unemployment tend to go hand in hand (Iyengar et al., 2007). Low basic mathematical skills have also shown to have an effect on opportunities for employment. A study in Great Britain (Gross, Hudson, & Price, 2009) found that low mathematical performance had undesirable employment outcomes for adults, even more than low reading skills, leading to lower rates of employment, lower rates of pay, and longer terms of unemployment. Similar findings have been found in studies completed in the United States (Geary, 2013).

Knowledge of the importance of academic skills and employment leads to the need for reading and mathematics interventions for postsecondary students with DD (Hua, Hendrickson, et al., 2012; Hua, Therrien, et al., 2012; Hua, Woods-Groves, Ford, & Nobles, 2014; Woods-Groves, Therrien, Hua, Hendrickson, Shaw, & Hughes, 2012). Students who continue to struggle with academics are often victims of poor early instruction and may be able to catch up, given appropriate academic interventions (Torgesen, 2005). With an emphasis on continued academic interventions in the postsecondary setting, there is a need to determine reliable and valid measures that can be used to measure the effectiveness of interventions. One possible tool for this is CBM

Developed in the late 1970s at the University of Minnesota's Institute for Research on Learning Disabilities (IRLD), CBM grew out of the idea that assessment and resulting instructional decisions based on those assessments should reference the curriculum used in the classroom (Deno, 1985). CBM is designed to be sensitive to growth in both general and special education settings, with quick administration and scoring leading to timely instructional decisions (Deno, 1985, 1992). Research has shown that the use of CBM measures in progress monitoring raises student achievement through responsive instructional decision making (Stecker, Fuchs, & Fuchs, 2005).

Existing research on CBM has largely been conducted at the elementary level (Foegen et al., 2007; Wayman, Wallace, Wiley, Tichá, & Espin, 2007). There is less research on CBM at the secondary level, but the few studies still support the strong research base for the use of CBM to monitor student progress and as an indicator of academic performance (Foegen et al., 2007; Johnson, Galow, & Allenger, 2013; Wayman, McMaster, Saenz, & Watson, 2010). This may also extend to the postsecondary level with recent research suggesting that CBMs can function as indicators of academic performance and growth for postsecondary students

with DD (Hosp, Hensley, Huddle, & Ford, 2014). Given the increasing number of postsecondary programs for students with DD and their emphasis on academic intervention and outcomes, this is an important step

In the aforementioned study, Hosp et al. (2014) analyzed CBM data for a group of postsecondary students with DD, all of whom attended a 2-year certificate program at a large university. Findings from the study are important as they provided preliminary support for criterion-related validity of CBM to function as an indicator of academic performance for postsecondary students with DD. Hosp et al. demonstrated that CBMs for reading, math, and writing display similar technical characteristics when used with postsecondary students with DD as with students at the elementary and secondary level. Despite the importance of these findings, the small sample size and the fact that no single study can conclusively demonstrate evidence of technical adequacy limit the implications and illustrate the need for direct replication.

Purpose

Given the importance of replication in establishing the robustness of research findings and the dearth of research on academic assessment with postsecondary students with DD, the following research question guided this study:

> **Research Question:** Can the criterion-related validity observed by Hosp et al. (2014) be directly replicated with different samples of postsecondary students with DD?

Method

Participants

Participants were students with cognitive/intellectual and DD (e.g., Autism) who enrolled at a 4-year public research institution in the Midwestern U.S. in a 2-year certificate program. The program is a hybrid program designed to facilitate young adults' independence and community integration. Students live on campus with others in the program as well as with typical college students. Program coursework includes instruction targeting career and independent life skills and integration in undergraduate courses with support. Academic coursework includes instruction in reading (e.g., increasing oral reading rate), mathematics (e.g., applied computation such as budgeting and tip calculation), and writing (e.g., editing skills). In addition to their coursework, students also participate in social activities offered by the university (e.g., sporting events, Dance-A-Thon).

Participants in this study included two cohorts ($N = 45$) of first semester postsecondary students with DD enrolled in the program described above. Data from Cohort A ($n = 24$) were collected in the fall of 2012. Participants were 41.6% female ($n = 10$) and 91.6% White ($n = 22$). The mean age of participants was 19.4 years ($SD = 1.10$). Data from Cohort B ($n = 21$) were collected in the fall of 2013. Participants were 47.6% female ($n = 10$) and 95.2% White ($n = 20$). The mean age of the participants was 19.3 years ($SD = 1.77$). All students had completed high school, earning a diploma or a certificate.

Instruments

We used CBMs from the aimsweb suite (Pearson Education, 2013) in this study. To replicate the study done by Hosp et al. (2014), Grade 4 materials were used for reading and Grade 5 materials were used for math.

OPR. OPR, also called Oral Reading Fluency (ORF) or Reading CBM (R-CBM), involves students reading aloud for 1 min from connected text. To calculate a score, the number of decoding errors a student makes is subtracted from the total number of words read. This metric is often referred to as words read correctly (WRC). Research involving students in Grades 1 to 6 has demonstrated OPR to have appropriate technical adequacy as an indicator of student reading skill, including comprehension. Validity coefficients for OPR have typically ranged from .60 to .80 and reliability coefficients from .82 to .99 (Reschly et al., 2009). Little research is available for older students. For students in Grade 10, validity coefficients have been found to range from –.02 to .71 with alternate-form and test–retest reliability observed at .91 (Espin & Deno, 1993a, 1993b). In a recent study involving postsecondary students with DD, Hosp et al. (2014) examined the technical characteristics of Grade 4 OPR with the *Woodcock–Johnson Tests of Achievement–Third Edition* (WJIII; Woodcock, McGrew, & Mather, 2001). Validity coefficients of .72 with Broad Reading, .63 with Reading Fluency, and .36 with Passage Comprehension were reported.

Maze. On Maze, students silently read a passage of 150 to 400 words. The first and last sentences remain intact while the rest of the passage has approximately every seventh word deleted. Where words are deleted, a choice of three words is provided. Students have to circle the word that best fits the sentence. For aimsweb Maze, a near distractor that is semantically correct and a far distractor that does not make sense in the passage are included along with the correct word (Shinn & Shinn, 2002). Performance on Maze is commonly measured by subtracting the number of incorrect restorations obtained by a student from the number of restorations attempted, thus creating the metric of correct restorations. For students of all age ranges, Maze validity coefficients have ranged from .60 to .86 while reliability coefficients have ranged from .68 to .90 (Wayman et al., 2007). Hosp et al. (2014) found validity coefficients with

Maze of .71 with Broad Reading, .66 with Reading Fluency, and .57 with Passage Comprehension compared with the WJIII.

Math Computation (M-COMP). CBMs for M-COMP include computational skills expected at the grade level they represent. M-COMP includes two pages of computational problems arranged in rows, and students are given 4 min to complete as many problems as possible. Performance on M-COMP is typically measured by the number of correct digits (CD) a student calculates. Reliability coefficients have ranged from .83 to .93 (Foegen et al., 2007), but research has mainly been limited to students in the elementary grades. Hosp et al. (2014) found validity coefficients for CD to be .67 for Broad Math, .75 for Math Computation, .73 for Math Fluency, and .40 for Applied Problems compared with the WJIII for postsecondary students with DD. In addition, Hosp et al. examined evidence of validity using the number of correct problems (CP) obtained by students. Validity coefficients for CP were observed to be .70 for Broad Math, .76 for Math Computation, .65 for Math Fluency, and .47 with Applied Problems.

Math Concepts and Applications (M-CAP). M-CAP measures applied mathematics skills (e.g., problem-solving). Students are given 30 math application problems and are told to complete as many possible in 8 min. Points (Pts.) are typically obtained per problem with more Pts. assigned to problems with multiple steps. Areas of mathematics included in M-CAP include number sense, operations, patterns and relations, probability, measurement, statistics, geometry, and algebra. For students in Grade 8, validity coefficients have been found to range from .61 to .87, while alternate-form reliability has ranged from .81 to .88 (Helwig, Anderson, & Tindal, 2002; Helwig & Tindal, 2002). While examining CP and Pts. with M-CAP, Hosp et al. (2014) found validity coefficients of .81 and .80, respectively, on the WJIII Broad Math cluster and .71 and .70, respectively, on Applied Problems. Correlations for Math Calculation (.77 and .74, respectively) and Math Fluency (.59 and .63, respectively) were less similar regarding their correlation.

WJIII. This study used the WJIII as a criterion measure. The WJIII is a standardized, norm-reference battery of achievement tests for use with individuals aged 2 to 90 years. The WJIII has three clusters (Broad Reading, Broad Math, and Broad Written Language) which have been observed to have strong reliability, generally at .90 or greater (Riverside, 2011). Each cluster is made of specific tests. The Broad Reading cluster includes Letter Word, Identification, Reading Fluency, and Passage Comprehension. The Broad Math cluster includes Math Calculation, Math Fluency, and Applied Problems. The Broad Written Language cluster includes Spelling and Writing Fluency but was not included in this study.

Procedures

All administration procedures were identical to those used in the Hosp et al. (2014) study. All CBM measures were administered during participants' regularly scheduled "special topics" course using the standardized aimsweb procedures. Group-administered measures (Maze, M-COMP, and M-CAP) were given in a single 40-min class session. All OPR passages were administered individually in a second 40-min class session the same week. Make-up administrations, for those who were late or absent for the first class, for group-administered measures and OPR were given individually during a "study table" session within 2 weeks of other CBM data being collected. Individuals responsible for administering the CBM measures included the study authors as well as faculty from program the participants were enrolled in and graduate students in special education. All administrators either had extensive experience in administering, scoring, and interpreting CBM measures or were provided an approximately 45-min-long training on administering and scoring CBM. WJIII data were independently collected by faculty from the postsecondary program.

Each CBM was scored by one of the study's authors. Overall, 20% of each scorer's probes were then blindly rescored by another author. Interscorer agreement was high for all five measures (99.9% OPR, 100.0% Maze, 98.4% M-COMP, and 98.3% M-CAP).

All student data were entered into a spreadsheet for analysis. All participants' data entries were reviewed to ensure accuracy regarding student results being correctly attributed to the student who obtained them. All entries were found to be entered accurately (i.e., 100%).

Data Analysis

Because the purpose of the study was a direct replication, we used the same data analysis procedures used by Hosp et al. (2014). Thus, we first calculated descriptive statistics for each CBM metric and the WJIII. Next, we calculated bivariate correlations between each CBM metric and each content-appropriate test from the WJIII (at the cluster and the individual level). Correlations were then compared using Meng's z to determine which metrics were statistically significantly better predictors (Meng, Rosenthal, & Rubin, 1992). Meng's z (Meng et al., 1992) is an accurate, simple method for comparing the relation between a dependent variable and multiple independent variables for the purpose of examining potential differences in how the independent variables predict performance on the dependent variable.

Table 1. Descriptive Statistics for Cohort A (n = 24).

Measure	Metric/cluster or test	*M*	*SD*	Skewness	Kurtosis	Range
OPR	WRC	132.08	62.29	1.41	4.60	36–341
Maze	CR	14.37	9.85	0.71	−0.01	2–36
M-COMP	CD	25.50	18.60	0.23	−0.89	0–64
	CP	8.50	6.80	0.80	1.38	0–28
M-CAP	CP	5.46	4.13	0.71	1.02	0–17
	Pts.	6.21	5.41	1.53	3.90	0–24
WJIII	Broad Reading	75.33	16.41	−1.29	1.55	30–97
	Reading Fluency	77.42	15.52	−0.25	0.12	41–110
	Passage Comprehension	77.17	18.22	−0.69	0.28	36–108
	Broad Math	62.79	25.93	−0.60	−0.11	8–109
	Calculation	68.63	27.83	−0.60	0.18	8–118
	Math Fluency	66.67	20.89	−0.12	−0.70	32–109
	Applied Problems	71.42	18.41	−0.61	−0.27	32–99

Note. OPR = Oral Passage Reading; WRC = words read correctly; CR = correct restorations; M-COMP = Math Computation; CD = correct digits; CP = correct problems; M-CAP = Math Concepts and Applications; Pts. = points; WJIII = *Woodcock–Johnson Tests of Achievement–Third Edition*.

Table 2. Descriptive Statistics for Cohort B (n = 21).

Measure	Metric/cluster or test	*M*	*SD*	Skewness	Kurtosis	Range
OPR	WRC	129.33	46.61	−0.27	−0.88	38–206
Maze	CR	16.95	11.44	1.06	0.84	2–46
M-COMP	CD	23.57	12.46	1.10	2.28	6–60
	CP	7.05	4.90	0.59	0.95	0–20
M-CAP	CP	5.19	4.03	0.81	−0.43	0–13
	Pts.	5.71	4.77	0.96	−0.20	0–16
WJIII	Broad Reading	77.90	17.28	−0.30	1.50	34–116
	Reading Fluency	76.95	11.60	−0.50	2.51	45–103
	Passage Comprehension	83.76	16.25	−0.03	0.64	46–116
	Broad Math	60.29	16.74	−0.52	−0.26	26–85
	Calculation	62.62	18.55	−0.47	−0.47	23–88
	Math Fluency	63.10	14.40	0.13	−1.00	37–88
	Applied Problems	70.43	14.44	0.02	−0.88	44–96

Note. OPR = Oral Passage Reading; WRC = words read correctly; CR = correct restorations; M-COMP = Math Computation; CD = correct digits; CP = correct problems; M-CAP = Math Concepts and Applications; Pts. = points; WJIII = *Woodcock–Johnson Tests of Achievement–Third Edition*.

To compare the samples, chi-square for categorical variables or t tests for continuous ones were conducted to demonstrate equivalence of the samples across demographic characteristics. In addition, a series of independent t tests was conducted to examine differences in students' academic performance on CBM measures and the WJIII. To address the replication aspect of our study, we compared performance from the original, published study (i.e., Hosp et al., 2014) with Cohort A as well as with Cohort B.

Examination of replication of the findings from Hosp et al. (2014) was conducted in two ways. First, the criterion-related validity evidence (i.e., the bivariate correlations between the CBM measures and metrics and the WJIII criterion measures) was compared using Rosenthal and Rosnow's (2007) formulas for tests of independent correlations. Second, the patterns of significant Meng's z were compared across all three cohorts (i.e., the original from Hosp et al., 2014 and the two replication cohorts: A and B) to identify replicated significant comparisons.

Results

Descriptive statistics for each CBM metric and the tests of from the WJIII can be found in Table 1 for Cohort A and Table 2 for Cohort B. Using guidelines provided by Tabachnick and Fidell (2013), with 1.0 being considered questionable and those greater than 2.0 to be problematic, each CBM metric was judged for deviation of skewness and kurtosis. Consistent with the findings of Hosp et al. (2014), most CBM metrics demonstrated acceptable levels of skewness and kurtosis. Issues with the distribution of our sample can be attributed to statistical outliers from each cohort

Table 3. Independent *t* Tests between CBM Measures and WJIII From Hosp, Hensley, Huddle, and Ford (2014) and Cohorts A and B.

		t(*p*)		
Measure	Metric/cluster or test	Hosp 2014/Cohort A (*df* = 63)	Hosp 2014/Cohort B (*df* = 60)	Cohorts A and B (*df* = 43)
OPR	WRC	0.431 (.668)	0.716 (.477)	0.166 (.869)
Maze	CR	0.158 (.875)	0.933 (.355)	0.813 (.421)
M-COMP	CD	1.886 (.064)	*2.318 (.024)*	0.403 (.689)
	CP	1.253 (.215)	*2.111 (.039)*	0.810 (.423)
M-CAP	CP	0.417 (.678)	0.643 (.523)	0.221 (.826)
	Pts.	0.510 (.612)	0.847 (.401)	0.327 (.746)
WJIII	Broad Reading	0.029 (.977)	0.680 (.499)	0.511 (.612)
	Reading Fluency	0.360 (.720)	0.248 (.805)	0.114 (.910)
	Passage Comprehension	0.959 (.341)	*2.793 (.007)*	1.273 (.210)
	Broad Math	0.050 (.960)	0.571 (.570)	0.378 (.707)
	Calculation	0.389 (.698)	1.697 (.095)	0.839 (.406)
	Math Fluency	0.643 (.523)	1.403 (.166)	0.658 (.514)
	Applied Problems	0.646 (.521)	0.439 (.662)	0.199 (.844)

Note. Values in italics are $p < .05$. CBM = curriculum-based measurement; WJIII = *Woodcock–Johnson Tests of Achievement–Third Edition*; OPR = Oral Passage Reading; WRC = words read correctly; CR = correct restorations; M-COMP = Math Computation; CD = correct digits; CP = correct problems; M-CAP = Math Concepts and Applications; Pts. = points.

(note range within Tables 1 and 2). Despite their effect on the distribution of our sample, we included outliers in the data for our analysis for two reasons. One, our sample size across cohorts is already small. Further reducing the number of students in each cohort would have further reduced the power of our analyses. Two, students with disabilities often perform on tests of academic skill significantly outside typical performance. Given the need for research with this population, it was determined prudent to keep all students in the database.

Comparison of Samples

Chi-square tests for gender comparisons between the samples were nonsignificant between the original and Cohort A, $\chi^2(2, 69) = 0.368$, $p = .544$, between the original and Cohort B, $\chi^2(2, 65) = 1.063$, $p = .303$, and between Cohorts A and B, $\chi^2(2, 45) = 0.161$, $p = .689$. *t* tests for age comparisons between the samples were significant between the original and Cohort A, $t(63) = 2.005$, $p = .049$, approaching significance between the original and Cohort B, $t(60) = 1.843$, $p = .070$, and nonsignificant between the cohorts, $t(43) = 0.231$, $p = .819$.

Results from conducting independent *t* tests related to reading can be found in Table 3. In regard to CBM reading metrics, there were no statistically significant differences between samples. The only comparison of a WJIII test that approached significance (.007 using a Bonferroni adjusted alpha of .003) was for Passage Comprehension when comparing the Hosp et al. (2014) sample with Cohort B, with Cohort B performing higher (83.76 to 73.57).

The independent *t* tests results for mathematics are also presented in Table 3. In regard to M-COMP metrics, there were no statistically significant differences when adjusting for multiple comparisons; however, both CD and CP were between .05 and .003 when comparing Hosp et al. (2014) with Cohort B, with Cohort B performing higher. No comparisons for M-CAP or the WJIII were statistically significantly different.

Reading

Correlations between reading-related CBM metrics and the WJIII, along with Meng's *z* results for examining potential differences in prediction for OPR versus Maze, can be viewed in Table 4 for both cohorts. A moderate relation for both OPR and Maze ($r = .67$ for both) was found with Broad Reading for Cohort A, whereas a strong relation ($r = .83$ for OPR; $r = .85$ for Maze) was found for Cohort B. Similar to the relation observed for Cohort B, Hosp et al. (2014) found a strong relation with Broad Reading. However, the relation for both measures found in their original study ($r = .72$ for OPR; $r = .71$ for Maze) appears to be practically more congruent with those of cohort A.

A moderate relation ($r = .62$) was found for OPR and Reading Fluency for Cohort A and a strong relation ($r = .73$) was found for Maze. For Cohort B, a strong relation was found for OPR and Maze ($r = .77$ for both) with Reading Fluency. Although the moderate relation found for OPR and Reading Fluency ($r = .63$) by Hosp et al. (2014) is not as strong as that observed for Cohort B, it is quite similar to our findings here in regard to for Cohort A. Furthermore,

Table 4. Correlations (and *p* Values) Between CBM Measures and WJIII From Hosp, Hensley, Huddle, and Ford (2014) and Cohorts A and B.

CBM	WJIII	Hosp et al. (2014)	Cohort A	Cohort B
OPR	Broad Reading	0.720	0.673 (<.001)[a]	0.827 (<.001)[b]
	Reading Fluency	0.632	0.619 (.001)[c]	0.771 (<.001)[d]
	Passage Comprehension	0.361	0.519 (.009)[e]	0.655 (.001)[f]
Maze	Broad Reading	0.714	0.666 (<.001)	0.851 (<.001)
	Reading Fluency	0.656	0.729 (<.001)	0.773 (<.001)
	Passage Comprehension	0.572	0.539 (.007)	0.689 (.001)
M-COMP CD	Broad Math	0.673	0.824 (<.001)	0.747 (<.001)
	Math Computation	0.752	0.815 (<.001)	0.722 (<.001)
	Math Fluency	0.726	0.753 (<.001)	0.718 (<.001)
	Applied Problems	0.397	0.780 (<.001)	0.533 (.013)
M-COMP CP	Broad Math	0.696	0.816 (<.001)	0.826 (<.001)
	Math Computation	0.764	0.791 (<.001)	0.813 (<.001)
	Math Fluency	0.648	0.781 (<.001)	0.740 (<.001)
	Applied Problems	0.467	0.776 (<.001)	0.609 (.003)
M-CAP CP	Broad Math	0.809	0.834 (<.001)	0.669 (.001)
	Math Computation	0.769	0.817 (<.001)	0.651 (.001)
	Math Fluency	0.592	0.704 (<.001)	0.549 (.010)
	Applied Problems	0.713	0.808 (<.001)	0.560 (.008)
M-CAP Pts.	Broad Math	0.800	0.816 (<.001)	0.642 (.002)
	Math Computation	0.737	0.798 (<.001)	0.631 (.002)
	Math Fluency	0.630	0.713 (<.001)	0.542 (.011)
	Applied Problems	0.699	0.784 (<.001)	0.540 (.011)

Note. CBM = curriculum-based measurement; WJIII = *Woodcock–Johnson Tests of Achievement–Third Edition*; OPR = Oral Passage Reading; M-COMP = Math Computation; CD = correct digits; CP = correct problems; M-CAP = Math Concepts and Applications; Pts. = points.
[a]$z = 0.076$, $p = ns$. [b]$z = -0.338$, $p = ns$. [c]$z = -1.211$, $p = ns$. [d]$z = -0.022$, $p = ns$. [e]$z = -0.187$, $p = ns$. [f]$z = -0.310$, $p = ns$.

the strong relations observed for Maze and Reading Fluency for Cohorts A and B suggest a stronger relation than that observed by Hosp et al. ($r = .66$).

A moderate relation was found for OPR for Cohort A ($r = .52$) and Cohort B ($r = .66$) for Passage Comprehension. A moderate relation was found for Maze for Cohort A ($r = .54$) for Passage Comprehension as well, whereas the relation for Cohort B approached being strong ($r = .69$). Findings from our current study suggest a stronger relation for OPR and Passage Comprehension compared with the weak relation ($r = .36$) found by Hosp et al. (2014). In addition, despite somewhat of a stronger relation between Maze and Passage Comprehension being observed for Cohort B, results of the current study appear congruent with the moderate relation ($r = .57$) found by Hosp et al.

Table 5 includes comparisons of the criterion-related validity coefficients in the Hosp et al. (2014) study and those found for Cohorts A and B. This yielded three sets of comparisons from our two replication samples: original to Cohort A, original to Cohort B, and Cohort A to Cohort B. Examination of the *z* scores and associated *p* values reveals that they are low with none approaching statistical significance (the smallest *p* value for OPR and Maze being .156). This indicates nonsignificant differences between the correlations found for the three samples.

For the Meng's *z* comparisons between OPR and Maze, Hosp et al. (2014) found no significant differences in prediction of the WJIII between the two CBM measures. The *z* results in Table 4 (as indicated by the superscripts a through f) also show no significant differences in prediction for OPR compared with Maze for Broad Reading, Reading Fluency, or Passage Comprehension on the WJIII for Cohorts A and B.

Mathematics

Correlations between mathematics-related CBM metrics and the WJIII can be found in Table 4 for Cohorts A and B. A strong relation was found for Broad Math and the M-COMP metrics for Cohort A ($r = .82$ for CD and CP) and Cohort B ($r = .75$ for CD; $r = .83$ for CP). These results are similar to the results observed by Hosp et al. (2014) for M-COMP CD and M-COMP CP ($r = .67$ and .70, respectively). The M-CAP metrics were found to also have a strong relation with Broad Math for Cohort A ($r = .82$ for CP; $r = .80$ for Pts.). This finding is congruent with the original Hosp et al. study ($r = .81$ for CP; $r = .80$ for Pts.). However, a moderate relation was found for the M-CAP metrics and Broad Math for Cohort B ($r = .67$ for CP; $r = .64$ for Pts.).

Table 5. Correlation *z* Test Comparisons Between Hosp, Hensley, Huddle, and Ford (2014) and Cohorts A and B.

		z(*p*)		
CBM	WJIII	Hosp 2014/Cohort A	Hosp 2014/Cohort B	Cohorts A and B
OPR	Broad Reading	0.34 (.737)	0.95 (.344)	1.13 (.259)
	Reading Fluency	0.08 (.937)	0.97 (.331)	0.93 (.351)
	Passage Comprehension	0.72 (.469)	1.42 (.156)	0.65 (.515)
Maze	Broad Reading	0.34 (.736)	1.27 (.203)	1.42 (.156)
	Reading Fluency	0.52 (.605)	0.85 (.398)	0.31 (.753)
	Passage Comprehension	0.18 (.861)	0.68 (.494)	0.76 (.449)
M-COMP CD	Broad Math	1.30 (.194)	0.52 (.600)	0.63 (.527)
	Math Computation	0.60 (.546)	0.23 (.818)	0.72 (.474)
	Math Fluency	0.22 (.826)	0.06 (.953)	0.24 (.812)
	Applied Problems	*2.30 (.022)*	0.61 (.543)	1.40 (.160)
M-COMP CP	Broad Math	1.05 (.294)	1.10 (.270)	0.10 (.924)
	Math Computation	0.25 (.802)	0.45 (.649)	0.19 (.848)
	Math Fluency	1.02 (.310)	0.62 (.532)	0.30 (.762)
	Applied Problems	1.95 (.052)	0.70 (.482)	1.02 (.307)
M-CAP CP	Broad Math	0.28 (.777)	1.10 (.271)	1.22 (.222)
	Math Computation	0.48 (.633)	0.84 (.400)	1.15 (.249)
	Math Fluency	0.72 (.475)	0.22 (.824)	0.80 (.421)
	Applied Problems	0.84 (.402)	0.91 (.363)	1.52 (.128)
M-CAP Pts.	Broad Math	0.17 (.865)	1.18 (.239)	1.19 (.233)
	Math Computation	0.55 (.583)	0.70 (.483)	1.09 (.276)
	Math Fluency	0.56 (.577)	0.47 (.639)	0.89 (.373)
	Applied Problems	0.70 (.484)	0.91 (.361)	1.41 (.160)

Note. Values in italics are $p < .05$. CBM = curriculum-based measurement; WJIII = *Woodcock–Johnson Tests of Achievement–Third Edition*; OPR = Oral Passage Reading; M-COMP = Math Computation; CD = correct digits; CP = correct problems; M-CAP = Math Concepts and Applications; Pts. = points.

A strong relation was found for Calculation and the M-COMP metrics for Cohort A ($r = .82$ for CD; $r = .79$ for CP) and Cohort B ($r = .72$ for CD; $r = .81$ for CP). Again, similar results were observed by Hosp et al. (2014) for M-COMP CD and M-COMP CP ($r = .75$ and .76, respectively). The M-CAP metrics were also found to have a strong relation with Calculation for Cohort A ($r = .82$ for CP; $r = .80$ for Pts.). Hosp et al. also observed a strong relation ($r = .77$ for CP; $r = .74$ for Pts.), but a moderate relation found for the M-CAP metrics and Calculation for Cohort B ($r = .65$ for CP; $r = .63$ for Pts.).

A strong relation was found for Fluency and the M-COMP metrics for Cohort A ($r = .75$ for CD; $r = .78$ for CP) and Cohort B ($r = .72$ for CD; $r = .74$ for CP). Once more, similar results were observed by Hosp et al. (2014) in their study ($r = .73$ for CD; $r = .65$ for CP). A strong relation was also observed for Cohort A for M-CAP metrics and Fluency ($r = .70$ for CP; $r = .71$ for Pts.). However, the moderate relation found for Cohort B ($r = .55$ for CP; $r = .54$ for Pts.) is more congruent with those found by Hosp et al. in their study ($r = .59$ for CP; $r = .63$ for Pts.).

A strong relation was found for Applied Problems and the M-COMP metrics for Cohort A ($r = .78$ for CD and CP), a noticeable difference from the Hosp et al. (2014) study which found a weak relation ($r = .40$ for CD; $r = .47$ for CP). In addition, a moderate relation was found for M-COMP metrics and Applied Problems for Cohort B ($r = .53$ for CD; $r = .61$ for CP). A strong relation was also observed for Cohort A for M-CAP metrics and Applied Problems ($r = .81$ for CP; $r = .78$ for Pts.). Hosp et al. also found a strong relation between the M-CAP metrics and Applied Problems ($r = .71$ for CP; $r = .70$ for Pts.); however, a moderate relation was found for Cohort B ($r = .56$ for CP; $r = .54$ for Pts.).

Table 5 includes comparisons of the criterion-related validity coefficients in the Hosp et al. (2014) study with those found for Cohorts A and B. This yielded three sets of comparisons from our two replication samples: original to Cohort A, original to Cohort B, and Cohort A to Cohort B. Examination of the *z* scores and associated *p* values reveals that they are low with only one reaching statistical significance level of $p < .05$ and none reaching statistical significance when adjusted for multiple comparisons (and a Bonferroni adjusted value of .003). This indicates nonsignificant differences between the correlations found for the three samples with only M-COMP CD to Applied Problems

Table 6. Correlations Between Mathematics CBM Metrics and WJIII Criterion Measures, Meng's *z*, Cohort A (*n* = 24).

	WJIII			
CBM metric	Broad Math	Math Calculation	Math Fluency	Applied Problems
M-COMP CD to M-COMP CP	0.226	0.650	−0.701	0.102
M-COMP CD to M-CAP CP	−0.147	−0.028	0.551	−0.367
M-COMP CP to M-CAP CP	−0.228	−0.303	0.755	−0.358
M-COMP CD to M-CAP Pts.	0.103	0.209	0.404	−0.046
M-COMP CP to M-CAP Pts.	0	−0.079	0.670	−0.086
M-CAP CP to M-CAP Pts.	0.470	0.474	−0.185	0.582

Note. CBM = curriculum-based measurement; WJIII = *Woodcock–Johnson Tests of Achievement–Third Edition*; M-COMP = Math Computation; CD = correct digits; CP = correct problems; M-CAP = Math Concepts and Applications; Pts. = Points.
*p < .1. **p < .05. ***p < .01.

Table 7. Correlations Between Mathematics CBM Metrics and WJIII Criterion Measures, Meng's *z*, Cohort B (*n* = 21).

	WJIII			
CBM metric	Broad math	Math calculation	Math fluency	Applied problems
M-COMP CD to M-COMP CP	−2.339**	−2.594***	−0.581	−1.667*
M-COMP CD to M-CAP CP	0.704	0.617	1.369	−0.197
M-COMP CP to M-CAP CP	1.521	1.513	1.522	0.358
M-COMP CD to M-CAP Pts.	0.988	0.830	1.515	−0.054
M-COMP CP to M-CAP Pts.	1.805*	1.735*	1.636	0.520
M-CAP CP to M-CAP Pts.	1.025	0.747	0.239	0.686

Note. CBM = curriculum-based measurement; WJIII = *Woodcock–Johnson Tests of Achievement–Third Edition*; M-COMP = Math Computation; CD = correct digits; CP = correct problems; M-CAP = Math Concepts and Applications; Pts. = Points.
*p < .1. **p < .05. ***p < .01.

correlations (p = .022) and M-COMP CP to Applied Problems correlations (p = .052) compared between Hosp et al. (2014) and Cohort A approaching the traditional significance criterion.

For the Meng's *z* comparisons for Cohort A, no difference in prediction for mathematics-related CBM metrics to performance on the WJIII was observed (see Table 6). Differences in prediction were observed for Cohort B (see Table 7), in favor of M-COMP CP compared with M-COMP CD, for Broad Math (z = −2.339, p < .05) and Math Calculation (z = −2.594, p < .01). As such, none of the findings from Hosp et al. (2014) regarding differences in prediction for mathematics-related CBM metrics to performance on the WJIII were replicated in either cohort as Hosp et al. found differences for Applied Problems when comparing M-COMP CD with M-CAP CP, M-COMP CP to M-CAP CP, and M-COMP CD to M-CAP Pts.

Discussion

The purpose of this study was to conduct a direct replication of Hosp et al. (2014) regarding the technical adequacy of CBM and postsecondary students with DD using two cohorts of similar students. Such a study is designed to demonstrate the importance of replication not only in intervention research but in assessment research and serve as a model of how it might be conducted. This purpose is consistent with improving research and the validity of results in all disciplines, including special education (Cook, 2014). Our results are mixed, consistent with research suggesting that study findings often fail to be replicated (Nosek, Spies, & Motyl, 2012; Pashler & Harris, 2012). However, direct replications, such as those reported here, are 4 times more likely to replicate findings than replications that contain "infidelities" which likely introduce additional sources of random error (Gilbert, King, Pettigrew, & Wilson, 2016). Perhaps this is as it must be as scholars, and the research community, seek to generate defensible knowledge.

Riley-Tillman and Burns (2009) discussed the importance of generating defensible knowledge in the context of experimentation and decision making in applied settings (i.e., are the interventions we are implementing in schools effective?), but we suggest such a process is equally important in issues related to measurement. In particular, it is important for researchers to generate defensible knowledge regarding which tools are appropriate for making instructional decisions about various student populations. Although a plethora of research exists demonstrating

the appropriateness of using CBM data for making instructional decisions about students in K-12 settings, such evidence is building in regard to postsecondary students with DD. The initial results of Hosp et al. (2014), though promising, must be further examined to generate defensible knowledge that such tools can be used with our population of interest.

The technical adequacy data obtained from the original Hosp et al. (2014) study, along with our two replications, support the claim that CBM can be used to make screening decisions about postsecondary students with intellectual disabilities and DD. For both replication cohorts, the criterion-related validity coefficients between the CBM measures and WJIII tests were moderate to strong for both reading and math. Overall, observed relations between CBM metrics and the WJIII are in the moderate to strong range (r = .52–.83) for OPR as well as for Maze (r = .54–.85). In addition, the overall observed relations between the M-COMP and M-CAP metrics and the WJIII (r = .53–.83 and r = .54–.83, respectively) were also in the moderate to strong range for our two replications. This replicates the findings from Hosp et al. and is further evidenced by the nonsignificant differences between the original correlations and those found for Cohorts A and B. It also provides a conceptual replication for the literature on criterion-related validity for CBM measures conducted with students in different grade levels and students who do not have intellectual disabilities or DD by finding similar relations with a different population.

Also, replicated from the original study was a lack of differences in criterion-related validity coefficients between OPR and Maze. This extends the findings from Hosp et al. (2014) by supporting the assertion that either reading measure can serve as an appropriate predictor of overall reading performance for postsecondary students with DD.

Where the findings were not replicated was with the Math CBM measures. Hosp et al. (2014) found differences in criterion-related validity between Math CBM measures and metrics when used to predict the Applied Problems test of the WJIII—which should be more closely aligned with M-CAP rather than M-COMP because both M-CAP and Applied Problems include skills beyond computation. In both replication cohorts of the present study, no differences were found in contrast to Hosp et al. Instead, significant differences were found between using CD and CP for M-COMP when used to predict Broad Math and Math Calculation for Cohort B. There were no differences for Cohort A. This indicates differences in magnitude of correlations among the original sample and both replication cohorts.

Looking for an explanation for this lack of replication within student performance or characteristics is difficult. Math performance for each group is not substantially different although a few differences between samples approached significance. Because the exact same procedures and measures were used for all three groups, it could be an indication that the Math CBM measures are not as robust as the reading ones. Recent changes in mathematics standards and instruction have deemphasized the use of computational algorithms and automaticity with fact recall (National Mathematics Advisory Panel, 2008). It could be that this has introduced a change in the underlying construct, and how it is being measured may not reflect it as accurately as it once did. It could also introduce increased variation in performance that is sample dependent to reflect the instruction those individuals had previously received in math as some states and districts adopt reforms sooner than others (Steiner-Khamsi, 2006).

Limitations

Although promising, our study's findings are not without limitations, and the results should be interpreted with caution. One limitation of our study is that despite participants representing several states and regions in the United States, the sample is neither nationally representative nor randomly selected. As such, inference of our results to all postsecondary students with DD may not be appropriate.

A second limitation, also related to the participants, is our small sample size in the two cohorts of students used to replicate the findings of Hosp et al. (2014). By separating students into their respective cohorts, we lost a degree of statistical power for the analyses we conducted. A larger sample size would certainly allow for greater ability to determine differences in correlations between CBMs and the WJIII. It is also possible that observed differences between the Hosp et al. study and our replications are due to the combined nature of the original study sample (i.e., first- and second-year students in the postsecondary program) versus the nature of our samples (i.e., both being first-year students only).

Another limitation is that, just as with previous research in this area, we only addressed evidence for criterion-related validity. Evidence for validity is multifaceted (see Shadish, Cook, & Campbell, 2002) and each category should be considered. The Joint Committee on Standards in Educational and Psychological Testing (2014) recognizes the need to examine multiple categories for evidence of validity as vital, stating all are required to determine the technical adequacy of a tool being used for assessment.

Implications and Future Research

Given the twice replicated finding of no difference in prediction for OPR and Maze to the WJIII, one could posit that there is a valid argument for using either to make screening decisions for postsecondary students with DD. Such a consistent finding also suggests that there is a valid argument

that a better tool is needed to measure the reading skills of postsecondary students with DD. Such a task will likely be a challenge given the observed strong correlation between the traditional reading-related CBM metrics examined and the WJIII. However, challenges in the literature as to the benefits of Maze exist (see January & Ardoin, 2012; Kendeou, Papadopoulos, & Spanoudis, 2012; Parker, Hasbrouck, & Tindal, 1992). Furthermore, the practice of reading aloud is an unnatural one for adult learners making OPR perhaps an unauthentic task for postsecondary students with DD. Other tools, such as the *Test of Silent Word Reading Fluency–Second Edition* (TOSWRF-2; Mather, Hammill, Allen, & Roberts, 2014) or the *Test of Silent Reading Efficiency and Comprehension* (TOSREC; Wagner, Torgesen, Rashotte, & Pearson, 2010) might prove to be viable alternatives. In addition, future research in this area may wish to examine methods for measuring silent reading skills using passages of connected text, neither the TOSWRF-2 (Mather, Hammill, Allen, & Roberts, 2014) nor the TOSREC (Wagner et al., 2010) use such an approach.

In regard to mathematics, although the overall coefficients were not statistically different, the patterns comparing different Math CBM measures with metrics within each sample were different. This variation in the significant differences between measures (i.e., M-COMP and M-CAP) and metrics (i.e., CD, CP, Pts.) presents a less clear picture of the utility of these measures with this population. However, results from our two replications provide evidence that there was no difference in the magnitude of predictions between the two replication samples, rather differences in the magnitude of predictions (as noted above) were observed comparing the original cohort with these two samples. Such a finding, again, points to the need for replication research in the field of special education. It also possibly suggests a need for future research to examine appropriate mathematics-related CBM tools for postsecondary students with DD.

The application of more rigorous research designs has been made for the study of students with disabilities in the postsecondary environment (Faggella-Luby, Lombardi, Lalor, & Dukes, 2014). Through replication of Hosp et al. (2014), a template for how such study may be done for examining the relation of CBM and postsecondary students with DD is provided. At this time, we think it prudent to focus research on this population as such students are often in need of instruction targeting basic academic skills, and CBM is a tool with the potential to be used for making decisions about students' response to instruction. Thus, postsecondary programs serving students with DD, which provide academic skill instruction, could improve student outcomes via appropriate use of CBM tools to guide instruction. Furthermore, we are aware of one study using CBM (i.e., Reading Rate) and postsecondary students (Lewandowski, Codding, Kleinmann, & Tucker, 2003). However, while Lewandowski et al. (2003) examined reading rate with a large population of postsecondary students ($N = 800$), students who self-disclosed a learning disability or difficulties with attention were specifically excluded from participation. Thus, it remains less clear that how CBM tools might be applied to the larger community of postsecondary students with disabilities attending traditional 2- and 4-year schools.

Conclusion

Replication is fundamental to the development of defensible knowledge in an evidence-based field such as education. Recent calls for replication in education have focused on intervention research because of its primacy as the heart of what we do (i.e., teach); however, assessment is also an important component of data-based decision making. We argue that assessment research is in just as much need of replications to develop a consistent evidence-based and ultimate standard of care. This study, through two direct replications of a study on criterion-related validity, offers a model for replication assessment research and highlights the need through both replicated and nonreplicated findings.

Authors' Note

The opinions expressed are those of the authors and so not represent views of the Office of Postsecondary Education (OPE) or the U.S. Department of Education.

Declaration of Conflicting Interests

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

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported herein was supported in part by the Office of Postsecondary Education (OPE), U.S. Department of Education, through Grant P407A100030 to the University of Iowa.

References

Cook, B. (2014). A call for examining replication and bias in special education research. *Remedial and Special Education*, *35*, 233–246.

Cook, B., Collins, L., Cook, S., & Cook, L. (2016). A replication by any other name: A systematic review of replicative intervention studies. *Remedial and Special Education*, *37*, 223–234.

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*. Independence, KY: Cengage Learning.

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

Espin, C. A., & Deno, S. L. (1993a). Content-specific and general reading disabilities of secondary-level students identification

and educational relevance. *The Journal of Special Education*, *27*, 321–337. doi:10.1177/002246699302700304

Espin, C. A., & Deno, S. L. (1993b). Performance in reading from content area text as an indicator of achievement. *Remedial and Special Education*, *14*, 47–59. doi:10.1177/074193259301400610

Education Sciences Reform Act of 2002, Pub. L. No 107-279, SS 112 (2002).

Faggella-Luby, M., Lombardi, A., Lalor, A. R., & Dukes, L., III. (2014). Methodological trends in disability and higher education research: Historical analysis of the journal of postsecondary education and disability. *Journal of Postsecondary Education and Disability*, *27*, 357–368.

Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics: A review of the literature. *The Journal of Special Education*, *41*, 121–139. doi:10.1177/00224669070410020101

Gall, M., Gall, J., & Borg, W. (2006). *Educational research: An introduction* (8th ed.). New York, NY: Pearson.

Geary, D. C. (2013). Early foundations for mathematics learning and their relations to learning disabilities. *Current Directions in Psychological Science*, *22*, 23–27.

Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children*, *71*, 149–164.

Gilbert, D., King, G., Pettigrew, S., & Wilson, T. (2016). Comment on "estimating the reproducibility of psychological science." *Science*, *351*, 1037.

Gross, J., Hudson, C., & Price, D. (2009). *The long term costs of numeracy difficulties. Every child a chance trust and KPMG*. East Sussex, UK: National Numeracy.

Hart, D., Grigal, M., Sax, C., Martinez, D., & Will, M. (2006). Postsecondary education options for students with intellectual disabilities. *Focus on Autism and Other Developmental Disabilities*, *25*, 134–150.

Helwig, R., Anderson, L., & Tindal, G. (2002). Using a concept-grounded, curriculum-based measure in mathematics to predict statewide test scores for middle school students with LD. *The Journal of Special Education*, *36*, 102–112. doi:10.1177/00224669020360020501

Helwig, R., & Tindal, G. (2002). Using general outcome measures in mathematics to measure adequate yearly progress as mandated by Title I. *Assessment for Effective Intervention*, *28*, 9–18. doi:10.1177/073724770202800102

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, *71*, 165–180.

Hosp, J. L., Hensley, K., Huddle, S. M., & Ford, J. W. (2014). Using curriculum-based measures with postsecondary students with intellectual and developmental disabilities. *Remedial and Special Education*, *35*, 247–357.

Hua, Y., Hendrickson, J. M., Therrien, W. J., Woods-Groves, S., Ries, P. S., & Shaw, J. J. (2012). Effects of combined reading and question generation on reading fluency and comprehension of three young adults with autism and intellectual disability. *Focus on Autism and Other Developmental Disabilities*, *27*, 135–146.

Hua, Y., Therrien, W. J., Hendrickson, J. M., Woods-Groves, S., Ries, P., & Shaw, J. (2012). Effects of combined repeated reading and question generation intervention on young adults with severe intellectual disabilities. *Education and Training in Autism and Developmental Disabilities*, *42*, 72–83.

Iyengar, S., Sullivan, S., Nicholas, B., Bradshaw, T., Rogowski, K., & Ball, D. (2007). *To read or not to read: A question of national consequence* (Research Division Report No. 47). National Endowment for the Arts.

January, S., & Ardoin, S. (2012). The impact of context and word type on students' maze task accuracy. *School Psychology Review*, *41*, 262–271.

Johnson, E., Semmelroth, C., Allison, J., & Fritsch, T. (2013). The technical properties of science content maze passages for middle school students. *Assessment for Effective Intervention*, *38*, 214–223. doi:10.1177/1534508413489337

Joint Committee on Standards in Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Kendeou, P., Papadopoulos, T., & Spanoudis, G. (2012). Processing demands of reading comprehension tests in young readers. *Learning and Instruction*, *22*, 354–367.

Kennedy, C. (2005). *Single case designs for educational research*. Boston, MA: Pearson.

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

Lewandowski, L. J., Codding, R. S., Kleinmann, A. E., & Tucker, K. L. (2003). Assessment of reading rate in postsecondary students. *Journal of Psychoeducational Assessment*, *21*, 134–144.

Liming, D., & Wolf, M. (2008). Job outlook by education. *Occupational Outlook Quarterly*, *52*(2), 2–29.

Makel, M., & Plucker, J. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, *43*, 304–316.

Makel, M., Plucker, J., Freeman, J., Lombardi, A., Simonsen, B., & Coyne, M. (2016). Replication of special education research: Necessary but far too rare. *Remedial and Special Education*, *37*, 205–212.

Mather, N., Hammill, D. D., Allen, E. A., & Roberts, R. (2004). *Test of Silent Word Reading Fluency*. Austin, TX: PRO-ED.

Mather, N., Hammill, D. D., Allen, E. A., & Roberts, R. (2014). *Test of Silent Word Reading Fluency—Second Edition*. Austin, TX: PRO-ED.

Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, *111*, 172–175. doi:10.1037/0033-2909.111.1.172

National Joint Committee on Learning Disabilities. (2008, June). *Adolescent literacy and older students with Learning Disabilities*. Retrieved from http://www.ldonline.org/?module=uploads&func=download&fileId=755

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the national mathematics advisory panel*. Washington, DC: U.S. Department of Education.

Neuman, S. (Chair). (2002, February). *Scientifically-based research*. Symposium conducted at the Office of Elementary

and Secondary Education, U.S. Department of Education, Washington DC. Retrieved from https://www2.ed.gov/nclb/methods/whatworks/research/page_pg2.html

No Child Left Behind Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*, 615–631.

Papay, C., & Bambara, L. (2011). Postsecondary education for transition-age students with intellectual and other developmental disabilities: A national survey. *Education and Training in Autism and Developmental Disabilities*, *46*(1), 78–93.

Parker, R., Hasbrouck, J., & Tindal, G. (1992). The maze as a classroom-based reading measure: Construction methods, reliability, and validity. *The Journal of Special Education*, *26*, 195–218.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*, 531–536.

Pearson Education. (2013). *AIMSweb*. San Antonio, TX: Author.

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

Riley-Tillman, T. C., & Burns, M. K. (2009). *Single case design for measuring response to educational intervention*. New York, NY: Guilford Press.

Riverside. (2011). *Woodcock-Johnson III normative update*. Itasca, IL: Author.

Rosenthal, R., & Rosnow, R. (2007). *Essentials of behavioral research: Methods and data analysis* (3rd ed.). New York, NY: McGraw-Hill.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*, 90–100.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York, NY: Cengage Learning.

Shinn, M. R., & Shinn, M. M. (2002). *AIMSweb training workbook: Administration and scoring of reading maze for use in general outcome measurement*. Eden Prairie, MN: Edformation.

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools*, *42*, 795–819.

Steiner-Khamsi, G. (2006). The economics of policy borrowing and lending: A study of late adopters. *Oxford Review of Education*, *32*, 665–678.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson.

Therrien, W., Mathews, H., Hirsch, S., & Solis, M. (2016). Progeny review: An alternative approach for examining the replication of intervention studies in special education. *Remedial and Special Education*, *37*, 235–243.

Think College. (2014). *Programs database*. Retrieved from https://thinkcollege.net/college-search

Torgesen, J. K. (2005). Recent discoveries from research on remedial interventions for children with dyslexia. In M. Snowling and C. Hulme (Eds.), *The science of reading: A handbook* (pp. 521-537). Hoboken, NJ: Wiley-Blackwell.

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (2010). *TOSREC: Test of Silent Reading Efficiency and Comprehension*. Austin, TX: PRO-ED.

Wayman, M. M., McMaster, K. L., Sáenz, L. M., & Watson, J. A. (2010). Using curriculum-based measurement to monitor secondary English language learners' responsiveness to peer-mediated reading instruction. *Reading & Writing Quarterly*, *26*, 308–332.

Wayman, M. M., Wallace, T., Wiley, H. I., Tichá, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education*, *41*, 85–120. doi:10.1177/00224669070410020401

Webb, L., & Metha, A. (2016). *Foundations of American education* (8th ed.). New York, NY: Pearson.

Whitehurst, G. (2003). *The institute for education sciences: New wine, new bottles, a presentation by IES director Grover (Russ) Whitehurst*. Retrieved from http://ies.ed.gov/director/speeches2003/04_22/2003_04_22.asp

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson tests of achievement*. Itasca, IL: Riverside.

Woods-Groves, S., Therrien, W., Hua, Y., Hendrickson, J., Shaw, J., & Hughes, C. (2012). Effectiveness of an essay writing strategy for post-secondary students with developmental disabilities. *Education and Training in Autism and Developmental Disabilities*, *47*, 210–222.