12-2017

Can Brief, Evidence-Based Measures Be Effective RTI Screens in Urban Schools? A Preliminary Study

Cheryl C. Durwin
Southern Connecticut State University, durwinc1@southernct.edu

Dina Moore
Southern Connecticut State University, moored14@southernct.edu

Deborah A. Carroll
Southern Connecticut State University, carrolld1@southernct.edu

Can Brief, Evidence-Based Measures Be Effective RTI Screens in Urban Schools?

A Preliminary Study

Cheryl C. Durwin, Dina Moore, and Deborah A. Carroll

Southern Connecticut State University

Poster presented at the annual meeting of the Northeastern Educational Research Association, Trumbull, CT, October 20, 2017. Correspondence may be sent to Cheryl C. Durwin at durwinc1@southernct.edu.

Abstract

Inefficient response-to-intervention (RTI) screening in urban schools where many students read below grade-level may under-identify students *needing* intervention or over-identify students, over-burdening a limited-resource system. In a first-grade sample from one urban school, we evaluated the classification validity of two research-based screening measures—the Test of Silent Reading Efficiency and Comprehension and the Word Test-3 (WT3) Synonym subtest—as alternatives to the school's screening measure, the Fountas & Pinnell Benchmark Assessment System (BAS). The WT3 yielded high classification accuracy in identifying students who were receiving intervention services, and it outperformed the BAS. Practical implications for RTI screening are discussed.

Keywords: response-to-intervention; screening; standardized tests; classification; literacy

Can Brief, Evidence-Based Measures Be Effective RTI Screens in Urban Schools?

A Preliminary Study

Schools in urban areas serving large populations of students from lower-socioeconomic backgrounds are faced with insufficient personnel and resources to serve all who need response-to-intervention for reading (RTI; Abbott et al., 2008). Universal screening is an essential component of RTI, in which schools assess all students to identify those needing additional intervention. Screening tools should: be reliable, valid and practical, accurately differentiate risk for later failure from no-risk, and have high consequential validity, yielding a positive net effect where students are not disadvantaged by the assessment process (Jenkins & Johnson, 2016; Messick, 1989). Efficient screening is critical in schools where 60-80% of students read *below* grade-level (compared to the typical 20-40%) because assessment and intervention resources become severely strained (Abbott & Wills, 2012). Inadequate screens can result in under-identification of students who *need* intervention or over-identification of those who do not, causing additional burden on already taxed resources. Providing schools with adequate screens can reduce these errors and assist schools in identifying appropriate educational accommodations.

Our research investigates the classification validity of two quick, objective, norm-referenced, reliable and valid assessments as alternatives to RTI screens used by schools. The district in which we conduct research recently switched their RTI assessment from the Developmental Reading Assessment, Second Edition (DRA-2) to the Fountas & Pinnell Benchmark Assessment System (BAS). On the BAS, students read passages and retell story details, and teachers rate their fluency and comprehension. Accuracy, fluency, and comprehension scores are used to determine students' instructional and independent reading

levels. The BAS has satisfactory reliability and validity evidence. It yields test-retest reliability of .86 from fall to spring administrations in a sample of second and third graders, weak convergent validity with Degrees of Reading Power (r = .44), and moderate convergent validity with the Slosson Oral Reading Test-Revised (r = .69; Klingbeil, Mccomas, Burns, & Helman, 2015). However, research indicates poor classification accuracy (54%-71%) of the BAS as a screening measure (Burns, 2014; Klingbeil et al., 2015). Additionally, its practical limitations include: subjective scoring, lack of national norms, and extensive time for teacher training, test administration, and interpretation and decision-making, all of which detracts from instruction.

Compared to the BAS, the TOSREC is practical as well as supported by adequate reliability and validity evidence. The TOSREC is a practical assessment, with four equivalent forms, group or individual administration, and scoring that is objective, quick, and yields a norm-referenced interpretation. The assessment also has adequate alternate-forms reliability (.86-.95), test-retest (after 2 months) with alternate-forms reliability (.81-.87), and inter-scorer reliability (values exceeding .99) across all forms and grades (Wagner, Torgesen, Rashotte, & Pearson, 2010). TOSREC scores from Grades 1-5 show strong concurrent and predictive correlations with oral reading fluency (average coefficient of .734; Wagner et al., 2010), and classification accuracy of 90% in predicting whether students met criterion on a state mastery test (Johnson, Pool, & Carter, 2011). Scores from Grades 6-8 yield strong correlations with measures of word recognition, passage comprehension, and silent reading fluency (.70 to .83; Wagner et al., 2010). In our previous research, TOSREC yielded high classification accuracy (85%) for predicting risk in second graders from one urban school, and the scores functioned as well as DRA-2 in distinguishing at-risk second graders receiving services from those not at-risk (Durwin, Moore, & Carroll, 2017a). TOSREC's classification results also were consistent with previous research

evaluating the validity of the TOSREC with a sample of students in Grades 1-5 (Johnson et al., 2011).

Like the TOSREC, The Word-Test-3 (WT3) is a practical assessment, with each subtest taking about 5 minutes to administer and with scoring that is objective, quick, and yields a norm-referenced interpretation. In our research we only use the Synonyms and Antonyms subtests that together take about 10 minutes. In the Synonyms subtest, examiners orally present 15 individual words and say, 'Tell me another word for… (angry, street, etc.).' In the Antonyms subtest, examiners orally present 15 individual words and say, 'What is the opposite of… (win, dark, cold)?' The examiner's manual for the WT3 reports adequate reliability and validity overall (Bowers, Huisingh, LoGuidice, & Orman, 2014). The average test-retest reliability for the Synonyms subtest for ages 6 to 7 is .79 and the average internal consistency for this age group is .76. Median inter-scorer agreement is 94%. The manual also reports content validity evidence, criterion-related validity (e.g., scores differentiating typically achieving students and those with language disorders), and minimized racial bias (Bower et al., 2014).

In the present study, we examine the classification validity of the TOSREC and WT3 Synonyms subtest in a first-grade sample from one urban school. As part of our larger reading intervention project, we administered these tests at the beginning of the school year to all students and obtained information from the school regarding which children were receiving intervention services. Our study sought to replicate the classification results of the TOSREC from our previous research with second graders in the same school (Durwin et al., 2017a) and to evaluate the screening validity of the WT3 Synonym test, as this was the first time using this test as part of our assessment battery.

Method

Participants

Participants were 29 first graders (mean age = 6.43, *SD* = .33) from two classrooms in an urban public school. The school population is primarily lower-socioeconomic, with 85.8% of students eligible for free/reduced lunch. The school used the September scores on the BAS and professional judgment to identify 14 students as at-risk (in need of intervention) at the beginning of the school year: 10 received our reading intervention but not school services, 3 received our intervention *and* school services (e.g., ESL, literacy, speech), and 1 received only special education services. The remaining 15 students were 'typically achieving.' Therefore, there were four children who received school services and 25 who did not (10 received our intervention and 15 were typically achieving).

Assessments

Table 1 provides a description of the TOSREC and WT3.

Procedure

Undergraduate research assistants individually administered tests in October/November and in April/May. TOSREC Form A was used as a pretest, and Form C as post-test. TOSREC and WT3 were administered on separate days and were introduced as "word games." It took about 5 minutes per child for TOSREC, and 5-7 minutes for WT3. Scores on the BAS were obtained from the school after all post-testing was completed.

Results

We used pretest scores of the WT3 Synonym subtest and the TOSREC to examine their sensitivity, specificity, and classification accuracy in differentiating children who receive school

services (n=4) and those who do not (n=25). Risk was defined as TOSREC standard scores of 89 and below (below-average for grade, per the test manual) and WT3 scores of 85 and below.

- Sensitivity: the screen accurately identifies individuals who fail a criterion test or outcome (i.e., those who receive services). Low sensitivity means the screen overlooks truly at-risk students (Johnson, Jenkins, Petscher, & Catts, 2009).
- Specificity: the screen accurately identifies individuals who pass the later criterion measure (i.e., those who do not receive services). Low specificity indicates over-identification of students as at risk who really are not (Johnson et al., 2009).
- Classification accuracy refers to accurate identification of true positives and true negatives.

Tables 2 and 3 show the classification statistics for the TOSREC and WT3 Synonym tests. Overall, the WT3 Synonyms subtest yielded better classification results than TOSREC.

- The WT3 had a sensitivity of 75% and a specificity of 96%, with an overall classification accuracy of 93.1%. The test resulted in only one false positive and one false negative. The "false positive" student had a TOSREC score 1 standard deviation below the mean, and a below-grade level reading score on the BAS. This student was identified for services at the end of the year. This would make classification of this student a 'true positive,' and it would improve WT3's sensitivity to 80%, specificity to 100%, and classification accuracy to 96.7%. The false negative was a student receiving ELL services.
- TOSREC's low specificity of 64% was due to 9 false positives. However, 7 of the 9 children were in our reading intervention; 6 of these 7 had below-grade level reading scores on the BAS. Based on TOSREC and BAS scores, these 6 students should have

been receiving services. If so, this would improve TOSREC's sensitivity to 80%, specificity to 84.2%, and classification accuracy to 82.8%, consistent with previous research (Durwin et al., 2017a; Johnson et al., 2011).

We also examined classification statistics for the September administration of the BAS. The BAS yields reading levels, which is ordinal data (compared to interval scale of standard scores), with letters A-D reflecting kindergarten-level performance and E-J representing Grade 1 performance. We used reading levels A-D to indicate risk for reading failure and E-J to indicate no-risk. Table 4 shows the classification statistics for the BAS.

- The BAS yielded a sensitivity of 100%. However, this may be a biased result because the test, in part, was used to decide which students received services.
- The specificity and classification accuracy of the BAS were unacceptable for a screening measure. For specificity, experts recommend minimum levels of 70% to 80% (Catts et al., 2009; Compton et al., 2006; Johnson et al., 2010; Kilgus et al., 2014). As shown in Table 4, the sensitivity was 40% and the classification accuracy was 48%.
- The low specificity is due to 15 students identified as "at-risk" by the measure who were not receiving school services. These would be considered false positives. However, the school asked us to provide our reading intervention to nine of the 15 students because they considered the students in need of additional reading remediation. It is not clear why these students were not receiving services (other than possibly limited school resources) because these nine students had below-grade BAS reading levels of A-D (2 A, 4 B, 2 C, 1 D) comparable to the four students receiving services. If these students received services, this would have improved the overall

classification accuracy of the BAS to 79% and the specificity to 62.5%, which is still well below the acceptable standards for specificity (Catts et al., 2009; Compton et al., 2006; Johnson et al., 2010; Kilgus et al., 2014).

Discussion

Our study evaluated the classification validity of two brief, evidence-based, standardized measures as alternative RTI screens. We compared the sensitivity, specificity, and classification accuracy of TOSREC and the Word Test-3 Synonyms subtest to the BAS, which the school used in their screening process to select students for intervention services. Even though the BAS was used by staff to select children for services, and for that reason yielded 100% sensitivity in selecting students at risk, it did not meet the standards of experts for specificity, resulting in a high rate of false positives.

The present study did not replicate our prior results indicating high classification validity of the TOSREC with second graders. However, TOSREC's classification statistics improved when taking into account the nine 'false positive' students who were receiving our reading intervention and whom the school considered at-risk. There are several possible explanations for the limitations in classification of the TOSREC. First, it is difficult to reliably assess reading comprehension at the beginning of first grade, especially with students from impoverished backgrounds. Many children could not read silently, as required by the test. Also, in this sample, children in the 'at-risk' group were not receiving services specifically for reading problems. For example, some received literacy services because they were classified as English Language Learners or had speech problems. This reflects a poor criterion variable with which to judge the adequacy of the reading assessment as a screening measure. Finally, sensitivity and specificity data will vary depending on the pre-determined cut-scores for establishing risk (Klingbeil et al.,

2015). As we discussed earlier, this school uses BAS scores along with professional judgment to select students for services. Therefore, there is no pre-determined cut score; as we documented, several students identified as false positives by the TOSREC had below-grade level BAS performance comparable to those who *were* receiving services. Moreover, our follow-up of the first-grade sample at the end of the school year indicated that many of the 'false positive' students who were not receiving services at the beginning of the school year were identified for services by the end of the year, and a few who received services became ineligible by the end of the year (Durwin, Moore, & Carroll, 2017b). Therefore, it is difficult to judge the classification validity against a moving target.

Compared to the TOSREC, the WT3 Synonym subtest scores at pretest yielded high classification validity. Correcting for the one 'false positive' student who was identified for services by the end of the year improved WT3's sensitivity to 80%, specificity to 100%, and classification accuracy to 96.7%. A simple measure of oral vocabulary may be able to accurately differentiate those receiving services not specifically related to reading (e.g., ELL, speech).

We acknowledge that our conclusions are limited by the small sample within a single school. Additional research with larger samples and at different grade levels is needed to further evaluate the classification validity of the brief, norm-referenced measures we used in our study. The fact that brief, practical standardized tests can yield classification statistics comparable to, and in some cases better than, less practical school-based tests with poor psychometric properties suggests that these briefer tests may be a promising alternative as RTI screens. Research indicates that using a screening battery yields better classification accuracy than a single measure (Foorman et al., 1998; Jenkins & O'Connor, 2002). In our own reading intervention research, we use a battery of empirically validated assessments that together take less than 30 minutes.

Schools may want to use brief, evidence-based assessments to provide a value-added judgment to RTI decision-making based on their lengthier reading tests. Alternatively, schools can use brief assessments as an initial screen and follow-up with lengthier tests when necessary.

Many children from lower-SES backgrounds begin school lacking reading readiness and do not catch up to peers without early, intensive intervention (Hart & Risley, 2003; NCES, 2013; Reardon, 2011; Reardon, Valentino, & Shores, 2002). Urban schools need valid and practical screening measures to identify and serve the students in most need.

References

Abbott, M., & Wills, H. (2012). Improving the upside-down response-to-intervention triangle with a systematic, effective elementary school reading team. *Preventing School Failure: Alternative Education for Children and Youth*, *56*(1), 37–46.

Abbott, M., Wills, H. P., Kamps, D., Greenwood, C. R., Kaufman, J., & Filingim, D. (2008). The process of implementing a reading and behavior three-tier model: A case study in a Midwest elementary school. In C. R. Greenwood, R. Horner, T. Kratochwill, & I. Oxaal (Eds.), *Elementary school-wide prevention models: Real models and real lessons learned* (pp. 215–265). New York, NY: Guilford Press.

Bowers, L., Huisingh, R., LoGuidice, C., & Orman, J. (2014). *The Word Test-3: Elementary examiner's manual*. Austin, TX: Pro-Ed.

Burns, M. (2014, May 12). *How valid are instructional level estimates from the Fountas and Pinnell Benchmark Assessment System?* Poster presented at the conference of the International Reading Association conference, New Orleans, LA.

Durwin, C., Moore, D., & Carroll, D. A. (2017a, April). *Using the TOSREC as an initial RTI screen: A practical alternative for urban schools*. Poster presented at the annual meeting of the American Educational Research Association, San Antonio, TX.

Durwin, C., Moore, D., & Carroll, D. A. (2017b, July). *Using the TOSREC as an initial RTI screen in urban schools: A replication*. Proposal submitted to the annual meeting of the American Educational Research Association for the annual conference in April 2018, New York, NY.

Foorman, B. R., Fletcher, J. M., Francis, D. J. Carlson, C. D., Chen, D., Mouzaki, A., et al. (1998). *Technical report: Texas Primary Reading Inventory* (1998 Edition). Houston:

Center for Academic and Reading Skills, University of Texas Health Science Center at Houston and University of Houston.

Hart, B., & Risley, T. R. (2003). The early catastrophe: The 30 million word gap by age 3. *American Educator, 27*(1), 4-9.

Jenkins, J. R., & O'Connor, R. E. (2002). Early identification and intervention for young children with reading/learning disabilities. In R. Bradley, L. Danielson, & D. P. Hallahan (Eds.), *Identification of learning disabilities: Research to practice* (pp. 99–150). Mawah, NJ: Erlbaum.

Jenkins, J. R., & Johnson, E. (2016). *Universal screening for reading problems: Why and how should we do this?* Retrieved on July 16, 2016 from: http://www.rtinetwork.org/essential/assessment/screening/readingproblems.

Johnson, E. S., Jenkins, J. R., Petscher, Y., & Catts, H. W. (2009). Screening for early identification and intervention: How accurate are existing tools and procedures in predicting first grade reading outcomes? *Learning Disabilities Research & Practice, 24*, 174-194.

Johnson, E. S., Pool, J. L., & Carter, D. R. (2011). Validity evidence for the Test of Silent Reading Efficiency and Comprehension (TOSREC). *Assessment for Effective Intervention, 37*(1), 50–57.

Klingbeil, D. A., Mccomas, J. J., Burns, M. K., & Helman, L. (2015). Comparison of predictive validity and diagnostic accuracy of screening measures of reading skills. *Psychology in the Schools, 52*(5), 500–514.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5–11.

National Center for Education Statistics (2013). *The Nation's Report Card:A First Look: 2013 Mathematics and Reading* (NCES 2014-451). Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

Reardon, S. F. (2011). The widening academic-achievement gap between the rich and the poor: New evidence and possible explanations. In G. Duncan and R. J. Murnane (Eds.), *Whither opportunity: Rising inequality, schools, and children's life chances*. New York: Russell Sage Foundation.

Reardon, S. F., Valentino, R. A., & Shores, K. A. (2002). Patterns of literacy among U.S. students. *The Future of Children, 22*(2), 17-37.

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (2010). *Test of Silent Reading Efficiency and Comprehension (TOSREC) examiner's manual.* Austin, TX: Pro-Ed.

Table 1

Description of Tests

	The Test of Silent Reading Efficiency and Comprehension (TOSREC)	**The Word Test-3 (WT3) Synonyms Subtest**
Administration	• Examinees are given 3 minutes to read sentences from a grade-level test booklet and decide whether each sentence is true or false (e.g., "A cow is an animal.").	• Examiners orally present 15 individual words and say 'Tell me another word for… (angry, street, etc.).'
Scoring	• Raw scores are converted to grade-based standard scores with a mean of 100 and a SD of 15. Percentiles are also available.	• Raw scores are converted to age-based standard scores with a mean of 100 and a standard deviation (SD) of 15. Percentiles are also available.

Table 2

TOSREC Classification of Students as Risk or No-Risk

TOSREC Categories	*Actual Risk Classification*		
	Receives School Services (Risk)	No School Services (No-Risk)	Total
Risk (0-89)	2[a]	9[b]	11
No-Risk (90 or above)	2[c]	16[d]	18
Total	4	25	29

Note: Sensitivity: 2/4= 50%; specificity: 16/25= 64%; classification accuracy: 18/29 = 62%.
[a] True positives. [b] False positives. [c] False negatives. [d] True negatives.

Table 3

Word Test-3 Synonym Subtest Classification of Students as Risk or No-Risk

Synonym Categories	*Actual Risk Classification*		
	Receives School Services (Risk)	No School Services (No-Risk)	Total
Risk (0-85)	3[a]	1[b]	4
No-Risk (86 or above)	1[c]	24[d]	25
Total	4	25	29

Note: Sensitivity: 3/4= 75%; specificity: 24/25= 96%; classification accuracy: 27/29=93.1%
[a] True positives. [b] False positives. [c] False negatives. [d] True negatives.

Table 4

Fountas & Pinnell Benchmark Assessment System (BAS) Classification of Students as Risk or No-Risk

BAS Categories	*Actual Risk Classification*		
	Receives School Services (Risk)	No School Services (No-Risk)	Total
Risk (Levels A-D)	4[a]	15[b]	19
No-Risk (Levels E-J)	0[c]	10[d]	10
Total	4	25	29

Note: Sensitivity: 4/4= 100%; specificity: 10/25= 40%; classification accuracy: 14/29=48%
[a] True positives. [b] False positives. [c] False negatives. [d] True negatives.