2018

Evaluating the TOSREC as a Brief RTI Screen for Early Struggling Readers in Urban Schools

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

Evaluating the TOSREC as a Brief RTI Screen for Early Struggling Readers in Urban Schools

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 18, 2018. Correspondence may be sent to Cheryl C. Durwin at durwinc1@southernct.edu.

Abstract

Efficient response-to-intervention (RTI) screening is critical in urban schools where 60-80% of students read below grade-level. We evaluated the classification validity of the Test of Silent Reading Efficiency and Comprehension (TOSREC) in identifying students receiving school services using two first-grade cohorts in the same urban school. TOSREC slightly under-identified students receiving school interventions; however, most "under-identified" students did not receive interventions for reading problems. TOSREC accurately identified those not-at-risk, but over-identified students who were receiving our research-based reading intervention while waiting for school services. Results suggest that schools need sufficient resources to provide appropriate services for all students needing intervention.

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

Evaluating the TOSREC as a Brief RTI Screen for Early Struggling Readers in Urban Schools

Schools in urban areas serving large student populations from lower-socioeconomic backgrounds are faced with insufficient personnel and resources to serve all who need response-to-intervention (RTI; Abbott et al., 2008). Screening to identify those needing intervention—an essential component of RTI—should be reliable, valid and practical, accurately predict risk, and have high consequential validity, ensuring that 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 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.

Our research investigates the classification validity of the Test of Silent Reading Efficiency and Comprehension (TOSREC) as an alternative to RTI screens used by schools. The district where we conduct research has used the Developmental Reading Assessment-2 (DRA-2) and, more recently, the Fountas & Pinnell Benchmark Assessment System (BAS). These tend to have poor classification accuracy and practical limitations, such as subjective scoring, no national norms, and extensive time for teacher training, test administration/scoring, and decision-making (Burns, 2014; Durwin, Moore, & Carroll, 2017a, 2017b; Klingbeil, Mccomas, Burns, & Helman, 2015). In contrast, TOSREC takes 3 minutes, contains four equivalent forms, utilizes quick, objective scoring, and yields norm-referenced scores. It has adequate reliability, strong concurrent and predictive correlations with oral reading fluency, and strong convergent validity with word recognition, passage comprehension, and silent reading fluency tests (Johnson, Pool, & Carter, 2011; Wagner, Torgesen, Rashotte, & Pearson, 2010).

Thus far, one published study has evaluated the classification validity of TOSREC and found high classification accuracy (90%) in identifying Grade 1-5 students who failed to achieve proficiency on a state mastery test (Johnson et al., 2011). In our research with second graders in one urban school, TOSREC yielded high classification accuracy (85%) for predicting risk for reading failure and functioned as well as the DRA-2 in distinguishing typically-achieving students from those receiving school services (Durwin et al., 2017a). Our study with first graders in the same urban school yielded a higher classification accuracy than the BAS (62% v. 48%), yet the classification accuracy was still poor due to nine false positives, seven of whom received our reading intervention while waiting for school services (Durwin et al., 2017b). Our present study investigates the classification validity of the TOSREC with first grade cohorts from two school years within the same urban school as part of our larger reading intervention project (see Moore, Durwin, & Carroll, 2018).

Method

Participants

Participants were 49 students from the same two first-grade classrooms within the same school, just from two different school years (2016-2017 and 2017-2018). The school is in an urban area and serves a large proportion of children from lower-socioeconomic backgrounds (85.8% of students in the school are eligible for free/reduced lunch). Table 1 provides demographic data by cohort year. Children in this sample were part of our larger reading intervention project (Moore et al., 2018). At the beginning of the school year after pretesting was completed on our test battery, the authors and school staff selected children for the reading intervention based on test data (ours and the school's) and on staff's professional judgment regarding which children needed reading intervention. Some children were assigned to receive

only our reading intervention at the staff's request because children were being monitored for risk, others received our reading intervention *and* school services, and a few received only school services (at the school's request). The remaining children were considered typically-achieving controls. The school used response-to-intervention (RTI) which involved continually monitoring the progress of children considered at-risk and making decisions regarding interventions and school services at various points throughout the school year. Because the number of students receiving school services varied over the school year, we used the children's status at the end of the school year to classify them as having received: (1) no services throughout the year, (2) *only* our reading intervention, (3) our reading intervention *and* school services, or (4) only school services during all or part of the year.

Test of Silent Reading Efficiency and Comprehension (TOSREC)

TOSREC is administered as part of our larger assessment battery which is given near the beginning and end of the school year. In TOSREC, examinees read sentences from a grade-level test booklet within a 3-minute time limit and decide whether sentences are true or false (e.g., "A cow is an animal."). Raw scores are converted to grade-based standard scores (M = 100, SD = 15).

Procedure

Research assistants individually-administered the TOSREC as part of our assessment battery in October/November as a pretest (Form A) and in April/May as a post-test (Form C). All tests were introduced as "word games." Testing on our battery occurred over a 3-week period from October to November and from April through May, but because each test is administered to all students before moving on to the next test in the battery, the administration of the TOSREC occurred over a few days for the entire sample. The same was true for other tests in our

assessment battery. Our reading intervention occurred from January through April (for the general timeline, see Durwin et al., 2017a, 2017b; Moore et al., 2018).

Results and Discussion

We used TOSREC pretest scores to examine the test's sensitivity, specificity, and classification accuracy (see Table 2 for descriptions) in differentiating children who received school services alone or with our intervention (n = 16) and those who did not (n = 33). Risk was defined as TOSREC standard scores of 89 and below (below-average for grade, per the test manual).

Table 3 showing TOSREC classification statistics reveals poor classification accuracy (71.4%) and poor sensitivity (62.5%).

- The sensitivity was lower than the 75% obtained with our second graders (Durwin et al., 2017a) and much lower than the recommended 90% criterion (Jenkins & Johnson, 2016). However, of the 6 false negatives, five students received "school services only" (2 English Language Learners; 3 special education) and one student received our intervention (at the school's request) along with special education. Because these students did not have specific reading problems, it would make sense that TOSREC identified them as having no-risk.
- TOSREC's specificity of 75.8% was within the minimum standard of 70-80% (Catts, Petscher, Schatschneider, Bridges, & Mendoza, 2009; Compton, Fuchs, Fuchs, & Bryant, 2006; Kilgus, Methe, Maggin, & Tomasula, 2014). However, it was much poorer than the 91% obtained with our second-grade sample (Durwin et al., 2017a). Importantly, of the 8 false positives, 6 children received our reading intervention only and had poor reading comprehension (M = 84.0, SD = 3.32). If the school had provided them with RTI,

TOSREC's classification validity would be adequate, with 72.7% sensitivity, 92.6% specificity, 83.67% classification accuracy, aligning more closely with previous research (Durwin et al., 2017a; Johnson et al., 2011).

Several factors may explain why the results failed to replicate our second-grade findings showing high classification validity (Durwin et al., 2017a). First, many first graders could not read silently, as required by the test, making it difficult to reliably assess reading comprehension with this sample. Also, our school services group contained many children without specific reading problems. This represents a poor criterion variable with which to judge the adequacy of a reading assessment as a screening measure. Finally, even when combining cohorts from two years, our small sample limits the generalizability of our conclusions.

The reliability, validity, and practicality of TOSREC make it promising alternative as a brief RTI screen. Future research with larger samples at different grade levels is needed to further evaluate its classification validity. Experts recommend a screening battery to yield 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 brief, empirically-validated assessments. Schools may want to use a brief assessment like TOSREC as part of their RTI approach to provide a value-added judgment to decisions based on their lengthier tests or use brief assessments as initial screens and follow-up with lengthier tests when necessary.

Many children from lower-socioeconomic backgrounds begin school lacking reading readiness and do not catch up to peers without early, intensive intervention (Hart & Risley, 2003; NCES, 2015; Reardon, 2011; Reardon, Valentino, & Shores, 2002). Schools in urban areas serving large populations of at-risk students need effective screening tools to accurately identify the students in most need of intervention.

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.

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.

Catts, H.W., Petscher, Y., Schatschneider, C., Bridges, M. S., & Mendoza, K. (2009). Floor effects associated with universal screening and their impact on the early identification of reading difficulties. *Journal of Learning Disabilities*, *42*, 162-176.

Compton, D. L., Fuchs, D., Fuchs, L. S., & Bryant, J. D. (2006). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology*, *98*, 394-409.

Durwin, C. C., Moore, D., & Carroll, D. A. (2017a, April 30). *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. C., Moore, D., & Carroll, D. A. (2017b, October 20). *Can brief, evidence-based measures be effective RTI screens in urban schools? A preliminary study*. Poster

presented at the annual conference of the Northeastern Educational Research Association, Trumbull, CT.

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., & 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.

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.

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.

Kilgus, S. P., Methe, S. A., Maggin, D. M., & Tomasula, J. L. (2014). Curriculum-based measurement of oral reading (R-CBM): A diagnostic test accuracy meta-analysis of evidence supporting use in universal screening. *Journal of School Psychology*, *52*, 377-405.

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.

Moore, D., Durwin, C., & Carroll, D. A. (2018, October 18). *Efficacy of a dialogic reading Intervention for struggling first-graders in urban schools*. Poster presented at the annual conference of the Northeastern Educational Research Association, Trumbull, CT.

National Center for Education Statistics (2015). *The Nation's Report Card: 2015 Mathematics & Reading Assessments.* Institute of Education Sciences, U.S. Department of Education, Washington, D.C. Retrieved on February 23, 2018 from: https://www.nationsreportcard.gov/reading_math_2015/#reading/acl?grade=4.

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

Demographic Data for First Grade Cohorts

	2016-2017 Cohort	**2017-2018 Cohort**
Classroom		
Teacher A	14	14
Teacher B	14	7
Gender (% Female)	46.4	47.6
Mean Age (yrs)	6.46 (.322)[a]	6.29 (.351)
TOSREC Pretest Score	99.11 (15.04) [a]	95.67 (14.78)
Group		
Typically-achieving control	15 (53.6%)	8 (38.1%)
Researcher intervention only	6 (21.4%)	4 (19.0%)
Researcher intervention with School Services	6 (21.4%)	4 (19.0%)
School Services Only	1 (3.6%)	5 (23.8%)

[a] Standard deviation (SD) in parentheses.

Table 2

Description of Sensitivity, Specificity, and Classification Accuracy Statistics

Statistic	**Description**	**Interpretation**

Sensitivity	• The screening measure accurately identifies individuals who fail a criterion test or outcome.
	• In our study, the screen will accurately identify students who are receiving school services.
		• Low sensitivity means:
		  ○ the screen overlooks truly at-risk students (Johnson, Jenkins, Petscher, & Catts, 2009).
		  ○ there is a high rate of false negatives.
		• Experts recommend adequate sensitivity should be 90% (Jenkins & Johnson, 2016).

Specificity	• The screening measure accurately identifies individuals who pass the criterion.
	• In our study, the scree will accurately identify typically-achieving (i.e., those who do not receive services).
		• Low specificity means:
		  ○ the screen over-identifies students as at risk who really are not (Johnson et al., 2009).
		  ○ there is a high rate of false positives.
		• Experts recommend a minimum specificity of 70%-80% (Catts et al., 2009; Compton et al., 2006; Johnson et al., 2011; Kilgus et al., 2014).

Classification Accuracy	• The screening tool accurately identifies true positives and true negatives.	

Table 3

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)	10[a]	8[b]	18
No-Risk (90 or above)	6[c]	25[d]	31
Total	16	33	49

Note: Sensitivity: 10/16 = 62.5%; specificity: 25/33= 75.8%; classification accuracy: 35/49 = 71.4%.
[a] True positives. [b] False positives. [c] False negatives. [d] True negatives.