FORUM PEDAGOGICZNE

PSYCHOLOGICAL, PEDAGOGICAL AND LINGUISTIC ASPECTS OF SCHOOL DIFFICULTIES

2024, 14, 2.1: 17–31
p-ISSN 2083-6325; e-ISSN 2449-7142
DOI http://doi.org/10.21697/fp.2024.2.1.2

DEBORAH K. REED[1]
University of Tennessee, United States
ORCID 0000-0003-0874-1412

S. RYAN HALL[2]
Georgia State University, United States
ORCID 0009-0009-2616-8470

DAVID E. HOUCHINS[3]
Georgia State University, United States
ORCID 0000-0002-2094-7686

Received: 21.05.2024; revised: 08.07.2024; accepted: 11.07.2024

HIGH-RISK STUDENTS TAKING LOW-STAKES ASSESSMENTS: DO THE DATA REFLECT ABILITY OR EFFORT?[4]

UCZNIOWIE WYSOKIEGO RYZYKA PODEJMUJĄ TESTY O NISKIEJ STAWCE – CZY DANE ODZWIERCIEDLAJĄ ZDOLNOŚCI CZY WYSIŁEK?

Abstract: This exploratory study examined whether test-taking effort (TTE) might be a concern in monthly low-stakes testing of juvenile offenders' ($n = 50$) reading abilities. Among the graphs of 10 randomly selected students' scores, 6 showed large fluctuations in performance from administration to administration, and another 2 showed precipitous declines across time. For the full sample, most of the average changes in scores from month-to-month far exceeded the standard error of measurement and equated to a 1- to 3-grade-level difference in how students' reading performance could be interpreted. These changes could be positive or negative and varied within and across students. Most of the average testing times were below the expected

[1] **Deborah K. Reed**, PhD; Professor and Director, Tennessee Reading Research Center, University of Tennessee. E-mail address: dreed20@utk.edu.

[2] **S. Ryan Hall**, PhD; Research Associate, Learning Sciences, Georgia State University. E-mail address: rhall@gsu.edu.

[3] **David E. Houchins**, PhD; Professor of Special Education, Learning Sciences, Georgia State University. E-mail address: dhouchins@gsu.edu.

[4] Acknowledgement: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant #R324A190145 to Georgia State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

minimum of 5 min, but total testing time generally was not correlated with scores. Given the response validity concerns, recommendations are made for supporting TTE.

Keywords: test-taking effort, low-stakes tests, reading, juvenile justice

Introduction

Interim reading assessment data is considered useful for forming small groups of adolescents with similar needs, identifying the skills of highest instructional priority for each group, and monitoring individual students' progress (Reed et al., 2012). Yet, the quality of an educator's decision making is contingent, in part, upon the quality of the data generated. Research examining data quality most often has focused on the measures' predictive validity to summative reading test outcomes or longitudinal outcomes (e.g., Ritchey et al., 2015) and classification accuracy for those at risk of not reading proficiently (Kilgus et al., 2014). Less often, studies have explored sources of construct irrelevant variance associated with the testing environment (Christ et al., 2013), testing directions (Colón et al., 2006; Reed et al., 2012), and examiner error (Cummings et al., 2014; Reed et al., 2019). Notably, exploration of construct irrelevant variance associated with interim reading assessments has been limited to external influences on scores without considering the potential that students' own test-taking effort (TTE) might be affecting the accuracy of estimating their reading abilities.

Test-Taking Effort on Low-Stakes Assessments

TTE has been associated with scores on group-administered assessments the test taker perceives as having no bearing on grades or educational status, also referred to as low-stakes tests (Penk, Schipolowski, 2015; Wise, DeMars, 2005, 2010). A common and reliable TTE metric is the amount of time it takes a tester to respond to an item, with rapid responding considered indicative of low effort (Silm et al., 2020). The potential for noncredible responding also has been a concern in individually-administered measures with low stakes (Erdodi et al., 2017). For example, using incentives to increase TTE was associated with significant change in testers' intelligence quotient scores (Duckworth et al., 2011). Furthermore, the effect of TTE has been stronger for boys, Black students, older students, and those of lower ability (Duckworth et al., 2011; Silm et al., 2020; Soland, 2018).

When comparing content area performance, study results suggest that large-scale reading scores are more impacted by TTE than math scores (Soland, 2018). Although findings suggest students' motivation can decrease over testing administrations from pre- to posttest (Finney et al., 2016), no known study has examined the extent to which TTE influences reading assessment data gathered monthly on adolescents exhibiting both achievement and emotional-behavioral risks. Monthly measurement is consistent with progress monitoring for adjusting instructional decisions to better meet the dynamic needs of adolescents with reading difficulties (Reed et al., 2012), but it can be considered low stakes because the scores do not influence students' grades or result in immediate reclassification of students. We refer to the population with combined achievement and behavioral challenges as high-risk students, who are found in greater concentrations in juvenile correctional facilities (Gagnon et al., 2009).

High-Risk Adolescents in Correctional Facilities

The number of U.S. youth committed to correctional facilities averages 36,000 students daily and disproportionately includes Black students as well as those with emotional-behavioral disorders (Puzzanchera et al., 2022). The average reading performance of incarcerated adolescents is reportedly several grades below their age-matched peers (Sanders et al., 2021). Unfortunately, the transfer of academic records for incoming students can be delayed, making it difficult to accurately place students in an educational program (Reed, 2018). Moreover, youth offenders often have histories of sporadic school attendance and exhibit a lack of connectedness to educators, lowering their motivation for school success (Reed, Wexler, 2014).

The combined challenges these high-risk youth pose place greater pressures on obtaining accurate reading performance data and leveraging the use of that data to maximize instructional opportunities both within facilities and when students re-enroll in their communities' schools. Thus, it is important to consider how juvenile offenders' tendency for academic disengagement might be manifested in TTE. Because low TTE can downwardly bias scores (Silm et al., 2020), its presence also might mean that the true reading abilities of juvenile offenders has been underestimated, which might affect programmatic decisions within facilities.

Purpose and Research Questions

The purpose of this exploratory case study was to examine the possibility that high-risk adolescents in juvenile justice facilities exhibit low TTE on low-stakes reading assessments that make it difficult to plan and deliver the most efficacious instruction while they are committed or plan for their transition to regular school upon their release. Our inquiry was guided by the research question: To what extent do monthly reading test data reveal anomalous variation in juvenile offenders' reading performance suggestive of a relationship with TTE?

Method

Participants in this case study represent a subsample of those from a larger efficacy trial of reading intervention for juvenile offenders. The students were drawn from three facilities (two housing males, *n* = 48; one housing females, *n* = 60) located in a Southeastern U.S. state. Among the 108 total students over one year of participation in the study, 71% were Black, 7% Hispanic, 16% White non-Hispanic, and 6% multiracial. By grade-level enrollment, 1% of students were in Grade 6, 3% Grade 7, 6% Grade 8, 34% Grade 9, 30% in Grade 10, 25% Grade 11, and 1% Grade 12. Based on age, our sample was an average 1 year older than the typical chronological grade placement. They were committed to the facility for a minimum of 6 months for a variety of offenses, with one male facility housing the highest risk youth in state confinement. Across the sample, 18% of students were classified with emotional-behavioral disorders, 9% with learning disabilities, and 6% with other health impairment (often indicating attention deficit disorder).

Measure

Capti. Students were administered a battery of assessments at intake and each month to ensure sufficient data over their varying lengths of stay. These tests had no bearing on students' placement in courses and did not affect students' grades in their courses. Because test takers' response time has demonstrated a more robust association with test performance than self-reported effort (Silm et al., 2020), we limited our exploration of TTE to the Reading Efficiency subtest of the computer-adaptive Capti assessment of reading ability, which automatically captures total testing time in minutes and seconds (Charmtech Labs, 2020). Although there are several subtests in Capti, only Reading Efficiency was administered monthly. This is a silent reading fluency and comprehension task that requires students to select from three options the correct word to complete deletions in sentences. The test includes two passages and, due to the computer-adaptive functioning, students may see between 32-41 items and are anticipated to spend between 5-9 min responding. In all grades, scores are scaled with a minimum of 190, maximum of 310, and mean of 250. Scores also are converted into a grade equivalency. Item response theory marginal reliabilities for the Reading Efficiency task reportedly range from 0.711 to 0.878 for Grades 8-12 (Sabatini et al., 2019).

Procedures. The monthly administration was proctored by trained research assistants, who pulled students (*M* = 8 at a time; range = 2 to 10 per assistant) from their courses and took them to a quiet room in the facility that was not in use at the time. Students were seated at individual computers. They were aware that the testing was being done for research purposes only, but they were encouraged by the proctor to do their best. Data were collected 9 times over the course of one school year. However, individual students may have participated in different numbers of administrations due to entry/exit dates, security issues, refusals, and scheduling conflicts. Therefore, only those students with a minimum of three testing administrations in one school year were retained in the dataset (*n* = 50; see Table 1 for *n*-size by grade).

Data Preparation and Analysis

The research team extracted scores and associated testing times from the Capti system. Research on TTE typically uses response time per item to filter out particularly rapid responses and improve the accuracy of the performance estimation (Silm et al., 2020; Soland, 2018). Given there was no extant literature on TTE among juvenile offenders taking a low-stakes monthly test, we did not intend to filter out responses. Therefore, we used only total testing time to explore whether TTE might be an issue warranting further research.

First, we randomly selected 10 students (20% of the dataset) and graphed their scores to look for directly observable variations from month to month. Scores from interim reading assessment data commonly exhibit small fluctuations across testing waves, but the data still tend to have a positive linear trend (Van Norman, Parker, 2016). Capti's Reading Efficiency task has a consistent scale at each grade (190 to 310; *M* = 250), and the standard error of measurement (SEM) would suggest scores from close administrations might fluctuate between 6 and 12 points in Grades 8-12 (Sabatini et al., 2019). Therefore, we looked for unexpected spikes and decrements in performance from month to month.

For all students in the dataset, we calculated the change in scale scores, grade equivalent scores, and testing times from month-to-month to determine the mean and change (computed with scores in absolute value), with attention to changes that exceeded the SEM for the grade level. Finally, we calculated the correlation between the monthly scale scores and testing times.

Results

Overall, the exploration of high-risk adolescents taking low-stakes interim reading assessments revealed the potential for TTE to introduce construct irrelevant variance. This could be visually discerned in plots of randomly selected students' data (see Figure 1). Eight of our 10 randomly selected students had trajectories that suggested TTE might be an issue. All of these students exhibited performance above and below the mean scale score (250). Students 2066, 2027, 2055, 1042, 2019, and 4004 exhibited great fluctuations in performance that suggest their TTE may have varied considerably from month to month. Students 2044 and 4029 primarily exhibited great declines from their initial performance, suggesting a gradual decline in TTE.

For comparison, the two sample plots with less extreme changes are presented in Figure 2. Students 2067 and 2045 had fairly stable performance with scores always below the mean, despite visible fluctuations. Because each student had at least one spike in performance, we also compared the score changes to the SEM for the students' grade levels.

Table 1 provides the average scale score, grade equivalency score, and minutes of testing time per Capti administration as well as the average change in students' scores and testing times between each testing point. All 50 students in the final dataset are represented and separated by grade level. As can be seen, most of the test-to-test changes in average scale score far exceeded the SEM. Although the group means did not always show as much variance from test to test, individual students within the group exhibited different patterns of score increases and decreases that, when averaged in absolute values, reveal a greater amount of change. Often this equated to a 1- to 3-grade-level difference in how students' reading performance could be interpreted.

To explore whether less visible fluctuations might still suggest some influence of TTE, we compared the month-to-month change in performance for the students in Figure 2 to the SEM. Student 2067's spike of 18 points at the third administration exceeded the Grade 8 SEM, as did the 13-point decline in the fourth administration. Similarly, student 2045's 11-point spike in the second administration (followed by a 10-point decline) and 8-point decline in the seventh administration (followed by a 10-point increase) also exceeded the Grade 9 SEM.

Correlations between the scale scores and minutes of testing times for each administration are displayed in Table 2. There was only one moderate and statistically significant correlation (i.e., administration 3). The mostly positive correlations reveal that longer testing times were associated with higher scale scores and vice versa. The two negative values at administrations 7 and 8 suggest that shorter testing times were associated with higher scores and vice versa. As shown in Table 1, the average testing times generally were brief (1 min 45 sec to 5 min 18 sec). Not only did the times tend to be shorter than the suggested 5-9 min for this computer-adaptive task (Charmtech Labs, 2020), but they also could vary by plus or minus 1-2 min from administration to administration.

Discussion

This exploratory study sought to determine whether TTE might be a concern in monthly low-stakes testing of juvenile offenders' reading abilities. Graphing a random sample of 10 students revealed that most (80%) had observable trajectories that raised concerns about response validity. Consistent with existing literature associating low TTE with score declination over repeated administrations of a test (Finney et al., 2016), two of the randomly selected students' graphs revealed mostly declining scores. However, six other randomly selected students exhibited great fluctuations in performance that exceeded reasonable expectations. To our knowledge, these anomalous patterns have not previously been documented in the literature but were the most common pattern among our sample of high-risk youth. Only two of the randomly selected students demonstrated fairly stable performance across administrations, but even their relatively smaller spikes and decrements in scores exceeded the SEM for their grade levels.

The high potential for TTE also was suggested by the magnitude of the full sample's average score changes from month-to-month. Although it is expected that students will make gradual improvement over time when monitored frequently (Van Norman, Parker, 2016), the scale and GE score changes suggest more than reasonable movement. Moreover, this could be positive or negative change at any given administration for any student. At some point, most—though not all—students performed both well above and well below the mean scale score of 250, but this was so inconsistent within and across students that the average scale score and grade equivalency were usually below the instrument's mean and the student's grade placement, respectively. Although a primary function of monthly testing is to guide teachers' decision making (Reed et al., 2012), data such as in the present study make it difficult to obtain realistic estimates of students' reading abilities or to plan appropriate instruction. This is a noted issue with response validity (Silm et al., 2020), especially with low-stakes tests (Penk, Schipolowksi, 2015; Wise, DeMars, 2005, 2010).

Response times are the most common approach to detecting a TTE issue (Silm et al., 2020). The total testing times in the present study suggested students were rapidly responding. The Reading Efficiency task is supposed to take students 5-9 min to complete, depending on the computer-adaptive functioning, but mean testing times were almost always less than 5 min. Nevertheless, there was only one statistically significant, moderate correlation between average testing time and average scale score (administration 3 r = .346), and two correlations were negative but weak and non-significant (administration 7 r = -.172; administration 8 r = -.101). Whereas response times in the TTE literature typically are determined on a per item basis for the purposes of filtering out items that might downwardly bias the score, we had only the total testing time for exploring the potential for TTE. It could be that the time is not as reflective of our sample's effort during a task that was not inherently of long duration, as compared to other standardized reading tests that can take 30 min to over an hour. In addition, Soland (2018) observed that rapid responding did not account for students who "have an item on the screen for 5 minutes, pay it no attention, and select an arbitrary response" (p. 322). Hence, additional research is needed to understand rapid responding on time-efficient progress monitoring tests.

Limitations and Directions for Future Research

As an exploratory study, we had a relatively small sample (n = 50) who were committed to facilities in a single U.S. state. We also had a high proportion of students who were Black, older, and enrolled an average 1 year below their typical grade placement—characteristics found in previous studies to be more prone to low TTE (Duckworth et al., 2011; Silm et al., 2020; Soland, 2018). Thus, our findings should not be generalized to non-incarcerated adolescents or be interpreted as questioning monthly progress monitoring more broadly. Rather, we were attempting to fill a gap in the literature on TTE with this specific population and to check the trustworthiness of our data before conducting a more rigorous examination of TTE.

To the latter point, we did not collect typical self-reports from students about their TTE in each administration because previous research has indicated it has a less robust association with performance than testing time (Silm et al., 2020). Thus, we do not know when students might have been exerting more or less effort during testing, which precludes us from evaluating whether TTE was contributing construct irrelevant variance to the higher or lower scores of students with the common pattern of large fluctuations. Future research is warranted to better understand TTE in these patterns of performance and determine whether scores are downwardly or upwardly biased. This could improve estimations of juvenile offenders' reading abilities.

Implications

Overall, results of the present study suggest that TTE is an issue among high-risk adolescents taking low-stakes reading tests. Because such assessments may be necessary to guide instructional decision-making in juvenile justice facilities that often are delayed in receiving students' academic records (Reed, 2018), it might be worthwhile to implement proactive supports for TTE. These could include having educators and peers explain the purpose of the tests to incoming students, providing tangible rewards, and informing students about their performance (Finney et al., 2016; Wise, DeMars, 2005). A meta-analysis of TTE interventions found that external incentives (ES = 0.21) and test relevance (ES = 0.27) were effective at improving low-stakes test performance (Rios, 2021). However, the review was of educational achievement tests that would be longer in duration and more comprehensive in coverage than the Reading Efficiency task administered in the present study, so further research is needed on the effectiveness of the recommendations when used for progress monitoring.

Nevertheless, given the multiple indications of spurious data in our sample, it seems reasonable to try preventing issues with TTE among high-risk adolescents. These youth already have poor educational histories and suffer a lack of connectedness and motivation (Reed, Wexler, 2014), so it is all the more important to have accurate information about their abilities to maximize learning opportunities while they are in a juvenile justice facility and prepare for their transition to schools in their communities upon release.

References

Charmtech Labs. (2020). *Capti Assess with ETS ReadBasix*. https://www.captivoice.com

Christ, T.J., Monaghen, B.D., Zopluoglu, C., & Van Norman, E.R. (2013). Curriculum-based measurement of oral reading: Evaluation of growth estimates derived with pre–post assessment methods. *Assessment for Effective Intervention, 38*(3), 139-153. https://doi.org/10.1177/1534508412456417

Colón, E.P., & Kranzler, J.H. (2006). Effect of instructions on curriculum-based measurement of reading. *Journal of Psychoeducational Assessment, 24*(4), 318–328. http://doi.org/10.1177/0734282906287830

Cummings, K.D., Biancarosa, G., Schaper, A., & Reed, D.K. (2014). Examiner error in curriculum-based measurement of oral reading. *Journal of School Psychology, 52*(4), 361-375. http://doi.org/10.1016/j.jsp.2014.05.007

Duckworth, A.L., Quinn, P.D., Lynam, D.R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences, 108*(19), 7716–7720. https://doi.org/10.1073/pnas.1018601108

Erdodi, L.A., Abeare, C.A., Lichtenstein, J.D., Tyson, B.T., Kucharski, B., Zuccato, B.G., & Roth, R.M. (2017). Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV) processing speed scores as measures of noncredible responding: The third generation of embedded performance validity indicators. *Psychological Assessment, 29*(2), 148–157. https://psycnet.apa.org/doi/10.1037/pas0000319

Finney, S.J., Sundre, D.L., Swain, M.S., & Williams, L.M. (2016). The validity of value--added estimates from low-stakes testing contexts: The impact of change in test-taking motivation and test consequences. *Educational Assessment, 21*(1), 60–87. https://doi.org/10.1080/10627197.2015.1127753

Gagnon, J.C., Barber, B.R., VanLoan, C., & Leone, P.E. (2009). Juvenile correctional schools: Characteristics and approaches to curriculum. *Education & Treatment of Children*, *32*(4), 673– 696. https://psycnet.apa.org/doi/10.1353/etc.0.0068

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*(4), 377-405. https://doi.org/10.1016/j.jsp.2014.06.002

Penk, C., & Schipolowski, S. (2015). Is it all about value? Bringing back the expectancy component to the assessment of test-taking motivation. *Learning and Individual Differences, 42*, 27–35. https://doi.org/10.1016/j.lindif.2015.08.002

Puzzanchera, C., Hockenberry, S., & Sickmund. (2022). *Youth and the juvenile justice system: 2022 national report.* National Center for Juvenile Justice.

Reed, D.K. (2018). Education for U.S. youth in secure care: The sum of the parts is not whole. In D. Gallard, K. Evans, & J. Millington (Eds.), *Children and their education in secure accommodation: Interdisciplinary perspectives of education, health, and youth justice* (pp. 117-127). Routledge Books.

Reed, D.K., & Petscher, Y. (2012). The influence of testing prompt and condition on middle school students' retell performance. *Reading Psychology, 33*(6), 562-585. http://doi.org/10.1080/02702711.2011.557333

Reed, D.K., Cummings, K.D., Schaper, A., *Lynn, D., & Biancarosa, G. (2019). Accuracy and reliability in identifying miscues during oral reading. *Reading and Writing: An Interdisciplinary Journal, 32*, 1009-1035. https://doi.org/10.1007/s11145-018-9899-5

Reed, D.K., & Wexler, J. (2014). "Our teachers…don't give us no help, no nothin'": Juvenile offenders' perceptions of academic support. *Residential Treatment for Children and Youth, 31*(3), 188-218. http://doi.org/10.1080/0886571X.2014.943568

Reed, D.K., Wexler, J., & Vaughn, S. (2012). *RTI for reading at the secondary level: Recommended literacy practices and remaining questions*. Guilford Press.

Rios, J. (2021). Improving test-taking effort in low-stakes group-based educational testing: A meta-analysis of interventions. *Applied Measurement in Education, 34*(2), 85-106. https://doi.org/10.1080/08957347.2021.1890741

Ritchey, K.D., Silverman, R.D., Schatschneider, C., & Speece, D.L. (2015). Prediction and stability of reading problems in middle childhood. *Journal of Learning Disabilities, 48*(3), 298-309. https://doi.org/10.1177/0022219413498116

Sabatini, J., Weeks, J., O'Reilly, T., Bruce, K., Steinburg, J., & Chao, S-F. (2019). SARA reading components tests, RISE forms: Technical adequacy and test design (3rd ed.). *ETS Research Report 19-36*. ETS.

Sanders, S., Jolivette, K., & Harris, C. (2021). Improving the reading comprehension skills of systems-involved youth: A preliminary investigation of an underserved population. *Learning Disabilities Research & Practice, 36*(3), 201-212. https://doi.org/10.1111/ldrp.12254

Silm, G., Pedaste, M., & Täht, K. (2020). The relationship between performance and test--taking effort when measured with self-report or time-based instruments: A meta--analytic review. *Educational Research Review, 31*(100335), 1-22. https://doi.org/10.1016/j.edurev.2020.100335

Soland, J. (2018). The achievement gap or the engagement gap? Investigating the sensitivity of gaps estimates to test motivation. *Applied Measurement in Education, 31*(4), 312–323. https://doi.org/10.1080/08957347.2018.1495213

Van Norman, E.R., & Parker, D.C. (2016). An evaluation of the linearity of curriculum--based measurement of oral reading (CBM-R) progress monitoring data: Idiographic considerations. *Learning Disabilities Research & Practice, 31*(4), 199-207. https://doi.org/10.1111/ldrp.12108

Wise, S.L., & DeMars, C.E. (2005). Examinee motivation in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1–18. https://doi.org/10.1207/s15326977ea1001_1

Wise, S.L., & DeMars, C.E. (2010). Examinee non-effort and the validity of program assessment results. *Educational Assessment, 15*(1), 27–41. https://doi.org/10.1080/10627191003673216

Table 1. Average Scale Scores, Grade Equivalencies, and Testing Times by Grade and Administration

Grade	Capti SEM	Scale Score *M* (SD)	Scale Score Change /*M*/ [b], c	Grade Equivalent Score *M*	Grade Equivalent Change /*M*/	Testing Time (min:sec) *M*	Testing Time Change /*M*/ c
Administration 1[a]							
8 (*n* = 2)	6.3	246.0 (33.9)		6.5 (5.0)		2:47 (2:18)	
9 (*n* = 19)	7.5	239.4 (20.7)		6.4 (2.7)		4:50 (2:50)	
10 (*n* = 21)	7.6	257.8 (26.2)		8.2 (2.9)		4:36 (1:40)	
11 (*n* = 8)	10.2	251.3 (23.8)		8.3 (3.1)		4:58 (1:40)	
Administration 2							
8 (*n* = 2)	6.3	243.0 (36.8)	3.0 (2.8)	6.5 (5.0)	0.0 (0.0)	3:06 (2:26)	0:18 (0:08)
9 (*n* = 19)	7.5	239.8 (21.2)	**14.3** (14.1)	6.5 (2.7)	1.8 (1.8)	4:33 (1:53)	1:34 (2:04)
10 (*n* = 21)	7.6	253.3 (23.3)	**14.8** (14.2)	8.0 (2.9)	1.3 (1.9)	4:36 (1:39)	1:16 (1:22)
11 (*n* = 8)	10.2	255.3 (25.5)	**13.8** (14.6)	8.5 (2.6)	1.0 (1.1)	5:10 (1:08)	1:34 (1:26)
Administration 3							
8 (*n* = 2)	6.3	253.0 (25.5)	**10.0** (11.3)	8.5 (2.1)	2.0 (2.8)	5:32 (1:19)	2:39 (3:27)
9 (*n* = 19)	7.5	233.3 (17.8)	**15.0** (15.2)	5.7 (2.9)	2.0 (2.2)	3:31 (1:53)	1:50 (1:41)
10 (*n* = 21)	7.6	250.0 (23.2)	**14.7** (11.0)	8.0 (3.1)	1.2 (1.6)	4:41 (1:13)	1:20 (1:10)
11 (*n* = 8)	10.2	246.4 (24.0)	**17.6** (11.2)	8.0 (3.4)	2.0 (2.1)	3:37 (1:49)	2:15 (1:16)
Administration 4							
8 (*n* = 2)	6.3	245.5 (33.2)	**7.5** (7.8)	6.5 (5.0)	2.0 (2.8)	2:46 (2:12)	2:46 (3:31)
9 (*n* = 14)	7.5	236.5 (20.8)	**13.9** (12.2)	6.3 (3.0)	1.6 (1.7)	4:30 (1:31)	1:33 (1:41)
10 (*n* = 14)	7.6	253.3 (23.8)	**17.8** (19.4)	7.8 (2.9)	3.0 (2.7)	5:03 (1:57)	1:35 (1:06)

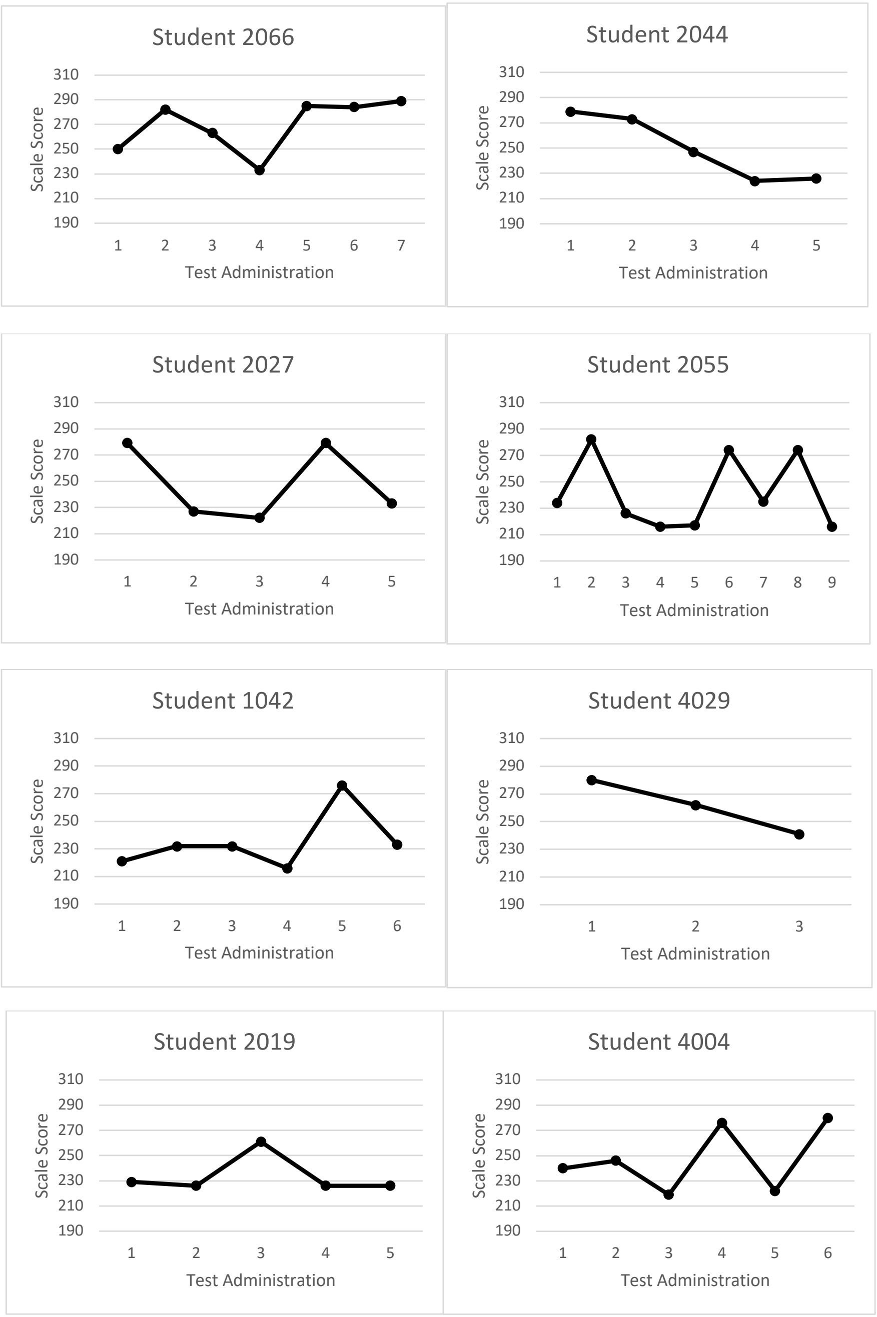
Grade	Capti SEM	Scale Score *M* (SD)	Scale Score Change /*M*/ [b], c	Grade Equi-valent Score *M*	Grade Equi-valent Change /*M*/	Testing Time (min:sec) *M*	Testing Time Change /*M*/ c
11 (*n* = 6)	10.2	242.0 (24.8)	**23.3** (9.1)	6.8 (3.5)	2.7 (2.3)	5:10 (2:59)	1:41 (2:26)
Administration 5							
9 (*n* = 14)	7.5	241.6 (22.2)	**12.6** (16.5)	6.0 (2.8)	1.4 (2.3)	3:29 (1:56)	1:34 (1:31)
10 (*n* = 10)	7.6	243.5 (23.6)	**17.0** (22.6)	6.8 (2.9)	2.4 (2.6)	4:42 (3:01)	2:50 (2:16)
11 (*n* = 6)	10.2	254.0 (33.8)	**13.3** (20.5)	7.3 (4.0)	0.8 (1.6)	3:39 (2:10)	1:38 (1:04)
Administration 6							
9 (*n* = 10)	7.5	238.2 (20.9)	**20.1** (17.2)	6.3 (3.1)	2.6 (2.4)	3:33 (2:07)	2:14 (1:12)
10 (*n* = 5)	7.6	261.2 (30.4)	**19.6** (24.8)	8.4 (3.4)	3.0 (2.4)	4:21 (2:04)	2:15 (2:23)
11 (*n* = 5)	10.2	264.4 (32.3)	**25.2** (28.3)	9.0 (3.5)	2.2 (3.3)	3:15 (1:26)	1:26 (1:10)
Administration 7							
9 (*n* = 8)	7.5	237.1 (20.3)	**10.4** (13.7)	5.9 (2.6)	1.3 (1.4)	5:25 (2:18)	1:56 (2:04)
10 (*n* = 4)	7.6	266.3 (22.8)	**16.3** (16.7)	9.0 (2.0)	2.0 (1.2)	3:30 (1:41)	0:37 (0:41)
11 (*n* = 5)	10.2	272.0 (19.3)	**23.6** (16.6)	10.6 (0.5)	2.0 (2.9)	3:34 (0:28)	1:03 (1:23)
Administration 8							
9 (*n* = 5)	7.5	248.2 (18.6)	**23.8** (12.5)	8.0 (2.3)	3.2 (1.8)	5:18 (2:02)	2:19 (1:26)
10 (*n* = 2)	7.6	282.5 (4.9)	3.5 (6.4)	10.5 (0.7)	0.5 (0.7)	3:01 (1:21)	0:25 (0:06)
11 (*n* = 4)	10.2	261.3 (33.0)	**11.5** (8.7)	8.3 (3.4)	2.3 (2.9)	3:08 (1:02)	0:58 (1:04)
Administration 9							
9 (*n* = 4)	7.5	223.8 (5.3)	**26.8** (25.3)	4.6 (1.9)	3.4 (2.9)	2:44 (1:02)	2:39 (1:56)
10 (*n* = 1)	7.6	285.0 (N/A)	1.0 (N/A)	11.0 (N/A)	0.0 (N/A)	2:04 (N/A)	0:00 (N/A)
11 (*n* = 2)	10.2	243.0 (22.6)	**15.5** (21.9)	6.0 (4.4)	1.3 (2.3)	1:45 (1:28)	1:21 1:40)

Note. [a]Administration number refers to the number of times a student was tested, but the date each student tested varied based on their time in the facility and other scheduling issues. [b]Change scores in bold exceeded the SEM for that grade level. cAll average change scores/times are computed only with the scores/times of the sample in consecutive waves. N/A = not applicable.

Table 2. Correlations of Reading Efficiency Scale Scores and Testing Times

Correlation	Admin 1	Admin 2	Admin 3	Admin 4	Admin 5	Admin 6	Admin 7	Admin 8	Admin 9
Reading Efficiency Scale Score and Testing Time	.140	.221	.346*	.049	.138	.337	-.172	-.101	.195

Note. * $p < .05$; Admin 1 = first test administration for a student (and so on with subsequent numbers).



Student 2066

Student 2044

Student 2027

Student 2055

Student 1042

Student 4029

Student 2019

Student 4004

Figure 1. Randomly Selected Students' Performance Across Test Administrations: Possible TTE

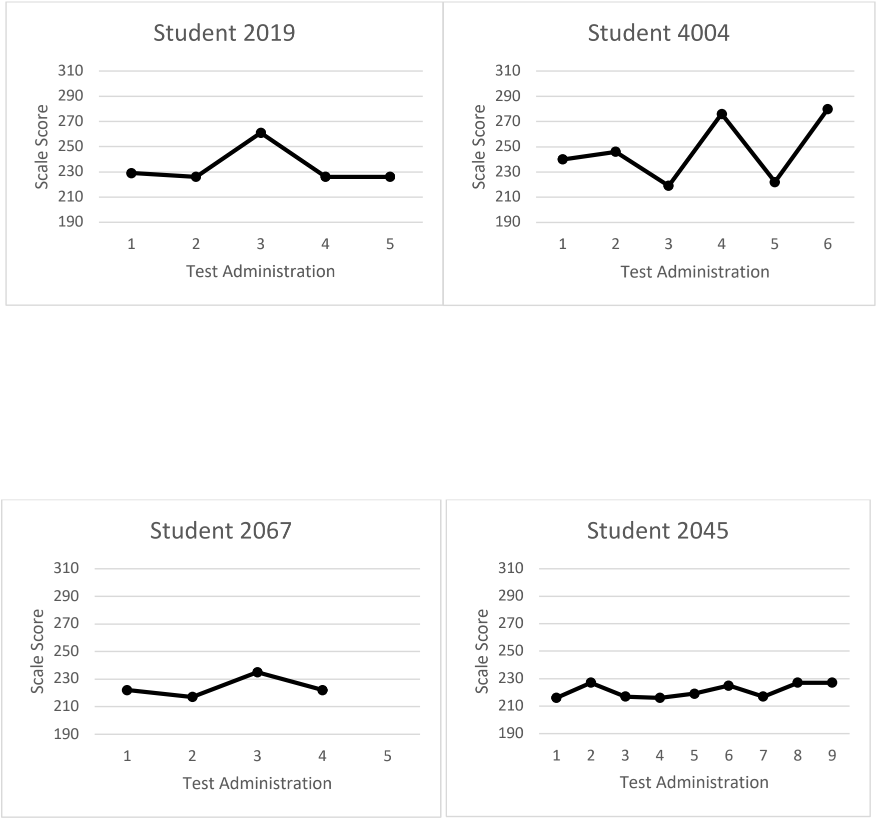


Figure 2. Randomly Selected Students' Performance Across Test Administrations: Less Likely TTE