Research Report

Comparing Test-Taking Behaviors of English Language Learners (ELLs) to Non-ELL Students: Use of Response Time in Measurement Comparability Research

ETS RR–21-25

Hongwen Guo
Kadriye Ercikan

December 2021

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Comparing Test-Taking Behaviors of English Language Learners (ELLs) to Non-ELL Students: Use of Response Time in Measurement Comparability Research

Hongwen Guo & Kadriye Ercikan

ETS, Princeton, NJ

In this report, we demonstrate use of differential response time (DRT) methodology, an extension of differential item functioning methodology, for examining differences in how students from different backgrounds engage with assessment tasks. We analyze response time data from a digitally delivered mathematics assessment to examine timing differences between English language learner (ELL) and non-ELL student groups. When matched on the total sum scores of the studied item form, results showed that ELLs spent a significantly longer time on most items compared to the non-ELLs who performed similarly on the test form. When matched on the total response time, results showed that ELL students spent a significantly longer time on items in the first half of the form but a shorter time on items in the second half. This research demonstrates the usefulness of DRT methodology in gaining insights about the differential engagement of students with assessment tasks.

Keywords Differential response time (DRT); measurement comparability; process data

doi:10.1002/ets2.12340

Digitally based assessments provide opportunities for capturing data that represent response processes like response time, sequences of actions, and tool uses when students respond to test questions. These response process data, referred to as process data in this report, may provide insights about test takers' behaviors, which in turn may provide valuable information relevant to comparing group performances and measurement comparability for these groups. Comparing group performances is central to the use and interpretation of assessment results. For example, national and state assessments are intended to provide information about students from different backgrounds to inform policy and practice, including students from different ethnic/racial backgrounds and English language learners (ELLs). Comparability of assessment results is central to the validity of interpretations and claims from such comparisons. This is particularly true when students from different language backgrounds, such as the ELL student groups, are compared to others. Research has shown that language load in assessments may lead to underestimation of knowledge skills and competencies for students who are taking tests in languages different than their home languages (Abedi & Herman, 2010; Ercikan et al., 2015). Test-taking behaviors can provide important insights about the degree to which students from different backgrounds have engaged with the assessment in similar ways. In addition, accommodations for ELLs on national and state assessments include extending testing time, providing bilingual dictionaries without definitions, translation, read aloud, small-group administration, and so on (National Center for Education Statistics, 2019). Research on whether ELLs behave differently from non-ELLs in terms of test-taking behaviors and time management may also shed light on whether accommodations work for ELLs.

In this study, we focused on item responses and timing data to compare differences between ELL and non-ELL students in mathematics performance and test-taking behaviors on a large-scale low-stakes assessment. We used statistical methods we proposed for examining differential response times in previous research (Ercikan et al., 2020; Guo & Ercikan, 2021) that used process data (such as item response time data and number of actions on an item) to evaluate differences in test-taking behaviors and effort among different language-and-cultural groups on a large-scale international assessment. Particularly, to analyze individual items, we used a methodology we proposed that is an extension of the widely used differential item functioning methodology (DIF; Dorans & Kulick, 1986; Dorans & Schmitt, 1991; Zwick & Thayer, 1996) for examining response time differences.

Corresponding author: H. Guo, E-mail: hguo@ets.org

The purpose of the current study was to demonstrate the use of differential response time (DRT) as an indicator of measurement comparability for investigating differences in response time behaviors between ELL and non-ELL test takers. In addition, we propose an effect size for response time difference, similar to the standardized mean difference (SMD; Zwick & Thayer, 1996), and a statistical test for significance. In the following two sections, we introduce the studied mathematics test and then describe the DIF and DRT methods. The next two sections present DIF and DRT results and some discussion of their possible implications for the studied assessment.

Data

The mathematics assessment used in this research measured students' knowledge and skills in mathematics and their ability to apply their mathematical knowledge in problem-solving situations. The assessment was a digitally based assessment administered in 2017 on a touch screen tablet with an attached keyboard. Some digitally based assessment questions included interactive on-screen tools (such as a ruler or scratch-work tool). Three formats or item types were used: multiple choice, short constructed response, and extended constructed response. Multiple choice items required students to read, reflect, or compute and then to select the alternative that best expressed the answer. Some items contained purely mathematical ideas and concepts, whereas others were set in the context of real-world problems. The number of items from each format type included in the assessment was selected so that the expected amount of testing time matched a prespecified amount.

The mathematics assessment consisted of a collection of mutually exclusive sets of items, referred to as forms. The forms were administered in separately timed 30-minute sessions. Each test taker in the assessment was administered two forms of items—a total of 60 minutes of testing. Several item forms were designated for public release after each assessment. In this study, we used one released form (referred to as Form 2) in the assessment for eighth graders to illustrate the proposed methods. The released data contained item responses and item response times. Item response time was the total time a student spent on solving the item and might include multiple visits to the item.

Assessment Form Information

There were 26,191 students who responded to the items on the assessment form, of whom 1,082 were ELL and 25,101 were non-ELL students. We removed students who spent a longer time than the allotted 30 minutes and those who had abnormal completions (such as multiple logins that led to inaccurately recorded timing data). There was a significant and large, approximately 1 *SD*, of the total score performance difference between ELLs and non-ELLs on the assessment form, with ELLs performing much lower (refer to Table 1). There was also a difference in the median response time between ELLs and non-ELLs, with ELLs spending a longer time in responding to the test questions. Note that the median response time on the form is the sum of item response times.

Item Information

The assessment form used in the research contained a total of 19 items. The total scores for the studied form ranged from 0 to 21. As shown in Table 2, among the items, 17 were scored as correct/incorrect and two items had partial credit scores (refer to the first column in Table 2). The second column in Table 2 shows the average item scores, which range from 0.22 to 1.52. Note that two items were polytomously scored, so the average item score could be larger than 1. In the third column of Table 2, the item discrimination index, item polyserial (Drasgow, 1986), is presented. Overall, the item discrimination statistics were relatively high, ranging from .41 to .76. The median response time, in the last column, shows that students took a longer time to respond to some items than to others, and the median item response time ranged from 22 to 120 seconds.

Methods

In this section, we describe the proposed DRT methodology and demonstrate its use for investigating differences in response time behaviors between ELL and non-ELL test takers. In a typical DIF analysis, students in the focal and reference groups are split into comparable subgroups based on a criterion/matching variable the test measures (usually

Table 1 Test and Sample Information

Test information	Non-ELL	ELL
Sample size	25,101	1,082
Total score mean (*SD*)	10.02 (4.74)	6.20 (3.54)
Total time median (IQR)	1,485 (514.25)	1,610 (516.75)

Note. Response time was recorded in seconds and ranged from 0 to 1,800 seconds. ELL = English language learner; IQR = interquartile range.

Table 2 Item Information

Item	Maximum score	Average score	Polyserial	Median RT (s)
1	1	0.56	0.50	22
2	1	0.67	0.61	59
3	1	0.52	0.55	70
4	1	0.82	0.64	106
5	1	0.36	0.76	101
6	1	0.35	0.41	47
7	2	1.52	0.57	120
8	1	0.22	0.42	77
9	1	0.22	0.5	54
10	1	0.72	0.68	67
11	1	0.4	0.59	60
12	2	0.87	0.76	82
13	1	0.25	0.74	92
14	1	0.33	0.65	62
15	1	0.63	0.75	40
16	1	0.39	0.51	23
17	1	0.57	0.62	26
18	1	0.35	0.59	46
19	1	0.68	0.60	48

Note. RT = response time.

proficiency or test sum score). A weighted sum of item score differences between comparable subgroups is computed as an effect size (Dorans & Kulick, 1986) to measure how differently the item functions for the two groups. The matching procedure attempts to reduce the impact of group proficiency differences by focusing on the performance of comparable members in the two groups on an item. A large positive (or negative) difference is typically taken as an indication that the studied item functioned differently for the two groups and members of the focal group scored higher (or lower) than comparable members of the reference group. In addition, statistical tests are available to test the significance of the difference (Zwick & Thayer, 1996). The same DIF methodology, often referred to as standardized P-DIF (STD P-DIF), was extended to study differential speededness for racial groups or gender groups (Dorans et al., 1988; Schmitt et al., 1991) on the old *SAT*® tests under the formula scoring instruction. However, most of the current tests use the right scoring instruction, and what was found on tests under formula scoring instruction may need to be reevaluated under the right scoring instruction with richer process data. In this study, in a similar fashion to the DIF methods, we extend the standardization DIF methodology to evaluate item DRT by comparing item response time differences between comparable members, with respect to either performance or overall response time, from the ELL and non-ELL groups. We propose an effect size for response time difference, similar to the SMD (Zwick & Thayer, 1996), and a statistical test for significance.

Performance and Response Time

The relationship between performance and response time is important in examining the degree to which differential response may be related to overall performance as well as the extent to which DRT might contribute to potential bias. To analyze the relationship between the total scores and the total response times, the locally weighted scatterplot smoothing

(LOESS) method (Cleveland et al., 1992), was used to regress the total form scores on the total form response times for ELL and non-ELL groups separately.

Compared to a parametric method, the LOESS method avoids a possible model misfit problem; compared to other nonparametric methods, the LOESS method is flexible and accurate. An additional advantage of LOESS is that it provides the confidence band of estimated regression curves (Cleveland et al., 1992).

Differential Response Time

As mentioned in the introduction, to compare response times on individual items between the matched ELL and non-ELL groups, we adapt the concepts of the STD P-DIF statistic (Dorans & Kulick, 1986) and the SMD (Zwick & Thayer, 1996). More formally, for a fixed item, each group is stratified into $(J + 1)$ subgroups based on the same criterion. Let T_{ijr} and T_{ijf} be response time of the studied item for student i in Stratum j in the reference group and in the focal group, respectively, where Stratum $j \in \{0, 1, 2, \cdots, J\}$; let n_{jr} and n_{jf} be the sample sizes in Stratum j for the two groups, and N_r and N_f be the total sample sizes of the reference and focal groups, respectively. In the following, we derive a statistical test for testing whether the average response time difference is significant between the two groups. For this, we need the following assumptions.

Because each test taker takes the test following their own pace, it is reasonable to assume the following:

Assumption 1. Given Stratum j, $\left\{ T_{ijg}, i = 1, 2, \cdots N; g = r, f \right\}$ are independent variables for student i and group g.

In the case that the strata are created by the test score, Assumption 1 indicates that given a student's performance on the test, the item response times are independent among test takers and between the focal and reference groups.

Assumption 2. Given Stratum j, T_{ijr}, and T_{ijf} follow the probability distribution with mean and standard deviation of $\left(\tau_{jr}, \sigma_{jr}\right)$ and $\left(\tau_{jf}, \sigma_{jf}\right)$, respectively.

This assumption indicates that students' item response times are a random sample of a distribution relevant to their score stratum and group membership.

When the above two assumptions hold, it is routine to show, by the central limit theorem, that, as min $\left(n_{j,f}, n_{j,r}\right) \to \infty$, in each stratum,

$$\sqrt{n_{j,f}}\frac{\widehat{\tau}_{j,r} - \tau_{j,f}}{\sigma_1} \Rightarrow N(0, 1) \text{ and } \sqrt{n_{j,r}}\frac{\widehat{\tau}_{j,r} - \tau_{j,r}}{\sigma_2} \Rightarrow N(0, 1), \tag{1}$$

where $\widehat{\tau}_{j,f}$ and $\widehat{\tau}_{j,r}$ are mean response time estimates in Stratum j for the focal and reference groups, respectively. Note that when σ is replaced by the sample standard deviation, the left-hand side of Equation 1 follows a t distribution, which can be well approximated by a normal distribution when the sample sizes are large.

Assumption 3. Across the strata j, $\widehat{\tau}_{jf} - \widehat{\tau}_{jr}$ are independent.

Therefore

$$\frac{1}{D}\sum_{j=0}^{J}\left[\sqrt{n_{j,f}}\frac{\widehat{\tau}_{j,f} - \tau_{j,f}}{\sigma_1} - \sqrt{n_{j,r}}\frac{\widehat{\tau}_{j,r} - \tau_{j,r}}{\sigma_2}\right]\frac{n_{j,f}}{N_f} \Rightarrow N(0, 1), \tag{2}$$

where

$$D^2 = \sum_{j=0}^{J}\left(\frac{n_{j,f}}{N_f}\right)^2.$$

Note that the weights $n_{j,f}/N_f$ in Equation 2 are the relative frequencies of the focal group, as was chosen for STD P-DIF (Dorans & Kulick, 1986). Also note that Assumption 3 may not hold in practice. However, if both focal and reference groups experience speededness, fatigue, or motivation issues in a similar way, Assumption 3 may be reasonably true. This reasoning applies to comparisons using either the total score or the total response time to stratify the groups. Therefore, under the null hypothesis that $\tau_{j,f} = \tau_{j,r}$ for $j \in \{0, \cdots, S\}$, the test statistic is

$$T = \frac{1}{\sqrt{2}D}\sum_{j=0}^{S}\frac{n_{j,f}}{N_f}\frac{\widehat{\tau}_{j,f} - \widehat{\tau}_{j,r}}{\sqrt{\widehat{\sigma}^2_{j,f}/n_{j,f} + \widehat{\sigma}^2_{j,r}/n_{j,r}}} \Rightarrow N(0, 1), \tag{3}$$

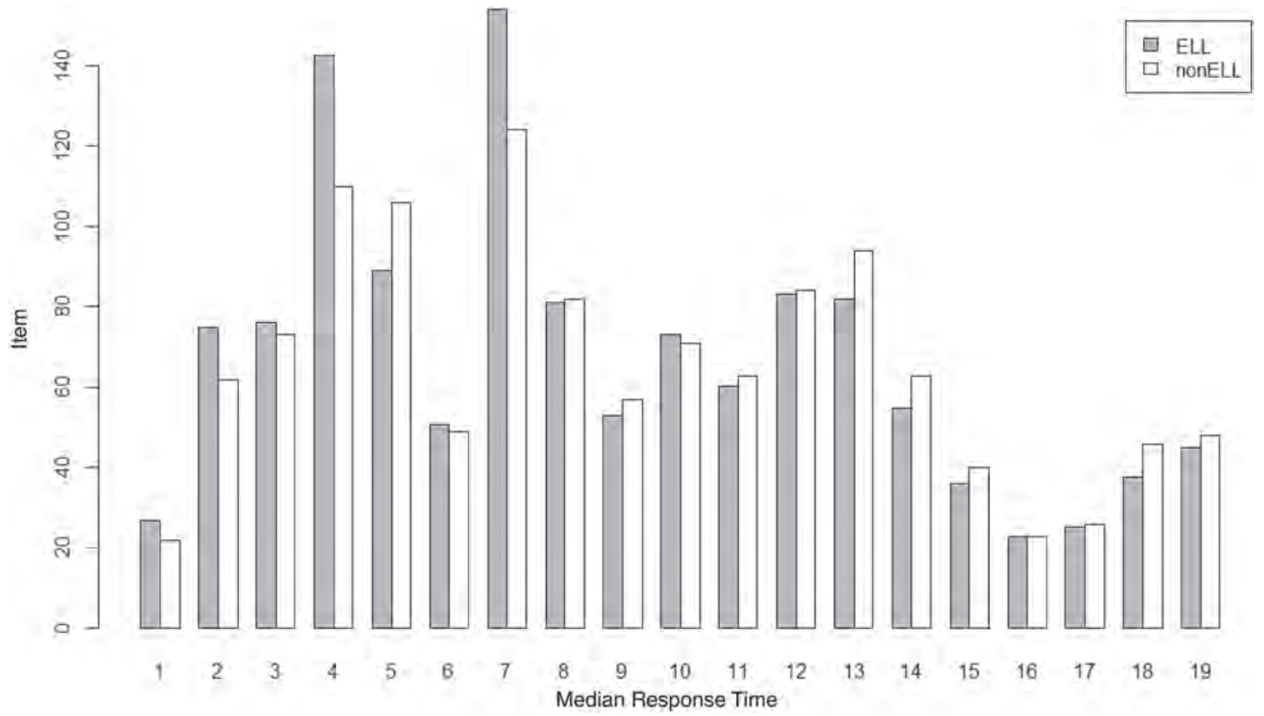

Figure 1 Median response times.

where $\widehat{\sigma}_{j,f}$ and $\widehat{\sigma}_{j,r}$ are consistent estimators of the standard deviation (say, the sample standard deviation). Note that in Equation 3, one can apply logarithm transformation of response time so that the test statistic converges to the standard normal distribution faster, because RTs tend to be positively skewed. Also note that, because response time is a continuous variable, this significance test is different from those for the SMD (Zwick & Thayer, 1996).

Like the STD P-DIF and the SMD, we also define the DRT effect size as

$$\mathrm{DRT} = \sum_{j=0}^{S} \frac{n_{j,f}}{N_f} \left[\widehat{\tau}_{j,f} - \widehat{\tau}_{j,r} \right]. \tag{4}$$

The DRT index measures the average difference between the conditional mean response times across strata between the focal and reference groups. When the groups are matched on (or stratified by) the total score, it is labeled as DRT_S, which shows, given similar proficiency in mathematics, whether one group spent a longer time on the studied item than the other group. On the other hand, when the groups are matched on (or stratified by) the total response time, it is labeled as DRT_T, which shows, given similar total response time, whether one group distributed a longer time on the studied item than the other group.

Results

Item Difficulty and Item Response Time

Figure 1 shows the bar plot of item median response time (y-axis) for each of the 19 items (x-axis). We observed that compared to the non-ELL group, the median response times were longer for ELLs in the first half of the items in the form but slightly shorter in the second half.

Bar plot of median response time for each item in the English language learner (ELL; shaded gray) and non−ELL groups (no shading).

Total Scores and Total Response Time

The form was administrated with a 30-minute time limit. Figure 2 shows the traced histograms of the total response time of Form 2 for the ELL group versus the non-ELL group. The traced histogram replaces bars by line segments connecting the

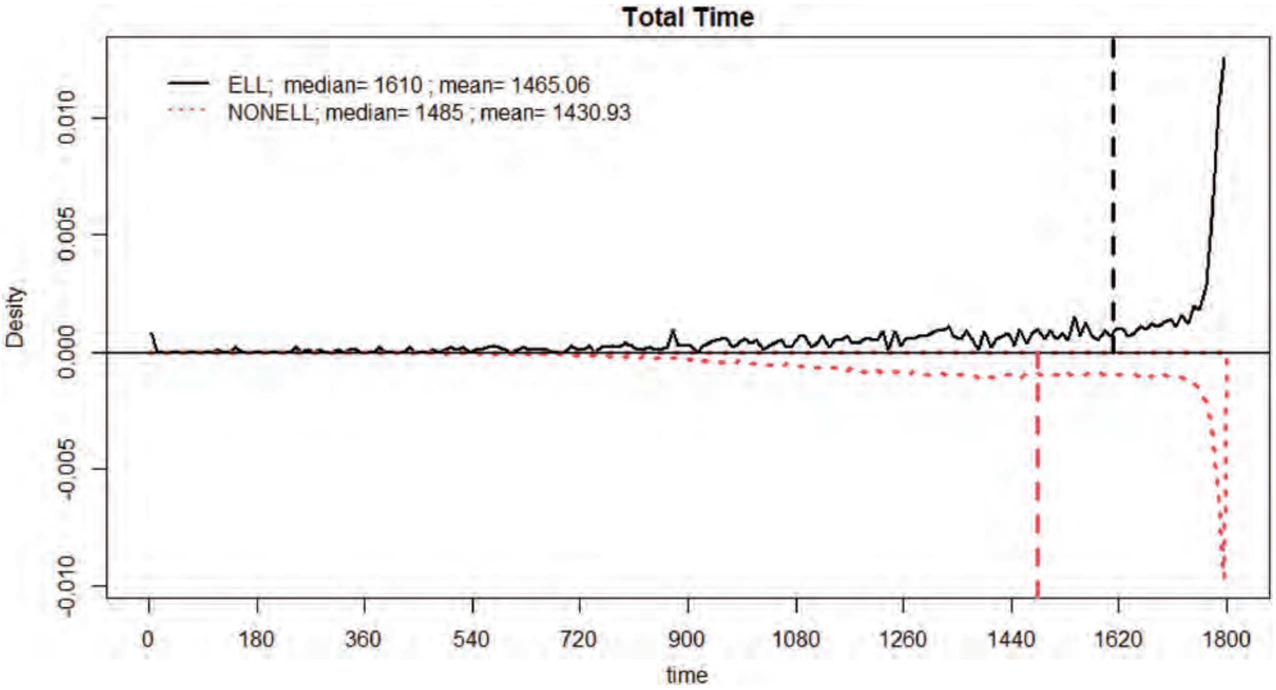

Figure 2 Traced histograms of the total response times for the English language learner (ELL) group (top, black) and the non–ELL group (bottom, red).

Table 3 Percentage of Test Takers Not Reaching the Item by English Language Learners (ELLs) and Non–ELLs

	Item position									
Group	≤10	11	12	13	14	15	16	17	18	19
ELL	3	4	5	7	10	13	15	17	18	19
Non-ELL	1	1	2	3	4	6	7	8	9	10

midpoints of bar tops. In the plot, the x-axis is time in seconds and the y-axis is the relative frequency. The ELL group had a slightly longer median response time than the non-ELL group (1,610 seconds vs. 1,485 seconds, labeled as the vertical dashed lines). The total time distributions are similar between the two groups: 36% of ELLs and 24% of non-ELLs ended the form in the last 1 minute.

Table 3 shows percentages of test takers who did not have a time stamp on certain items in the ELL and non-ELL groups. Overall, slightly more ELLs did not reach the later items in the form than non-ELLs. About 81% of ELLs and 90% of non-ELLs reached all the items in the form.

Figure 3 shows the traced histograms of the total score distributions of the ELL group versus the non-ELL group. The distributions are skewed to the right, indicating that few students obtained high scores on the form. In Figure 3, the average score was 6.20 for the ELL group, and that for the non-ELL group was 10.02. The median score was 6 for the ELL group and 10 for non-ELL group (labeled as the vertical dashed lines in the plot). These low scores indicate that this mathematics form was hard for the sampled eighth graders and that it was particularly hard for the ELL group.

Figure 4 shows the total response time (mean, median, interquartile range) comparison between the ELL and non-ELL groups for each score group. For each score group, ELL test takers tended to spend a slightly greater amount of time on the form than did non-ELL test takers. The light gray area is the interquartile range of the non-ELL group, and the blue hatched area is that of the ELL group. The score subgroups did not show clear differences in the total testing time, which explained the low correlation between scores and testing times. Note that no ELL students scored higher than 19. In addition, the response time distributions are mostly symmetric (i.e., median is close to mean) for each score subgroup.

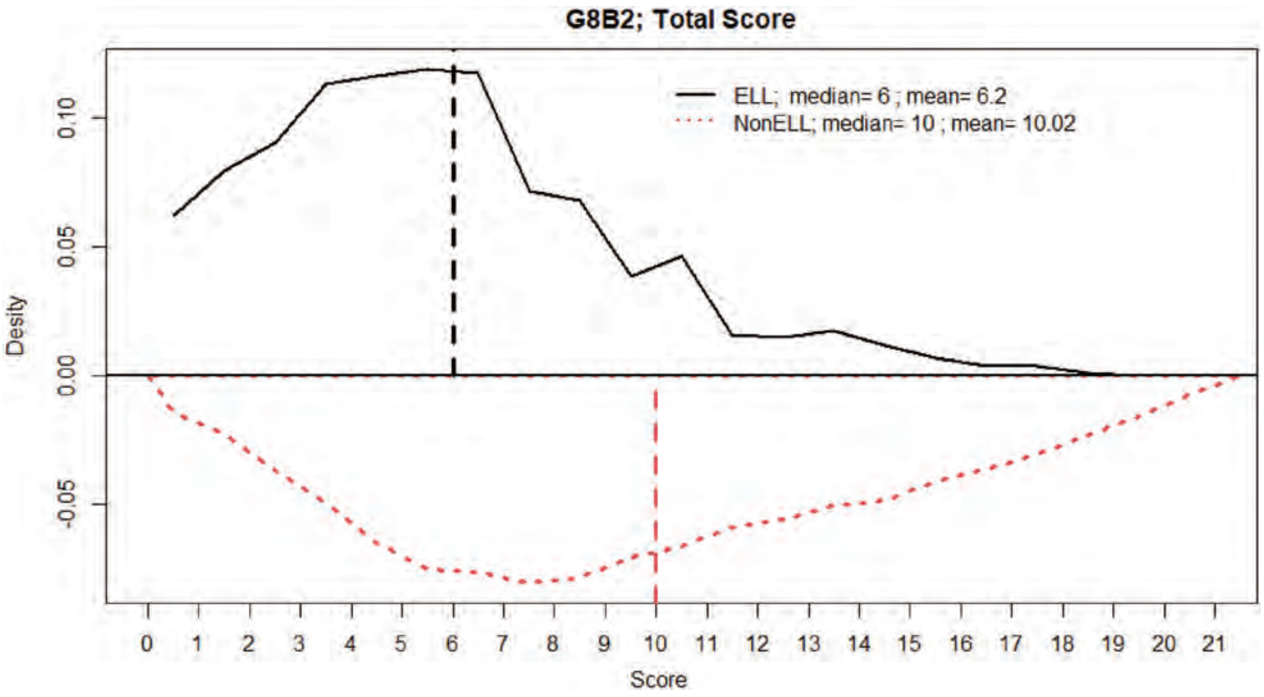

Figure 3 Traced histograms of form scores with a possible score range from 0 to 21 for the English language learner (ELL) group (top, black) and the non-ELL group (bottom, red).

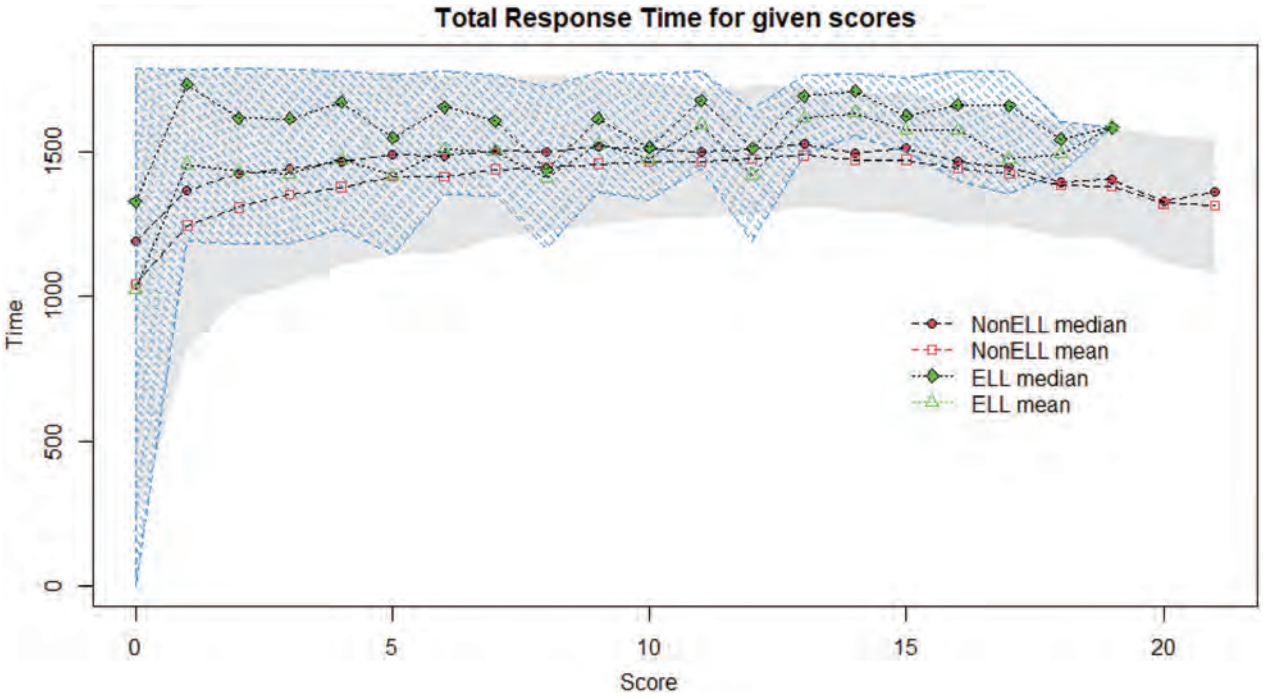

Figure 4 Total response time given total scores on Form 2 for the English language learner (ELL) and non-ELL groups. The gray area is the interquartile range of total response times for the non-ELL group, and the blue hatched area is that for the ELL group.

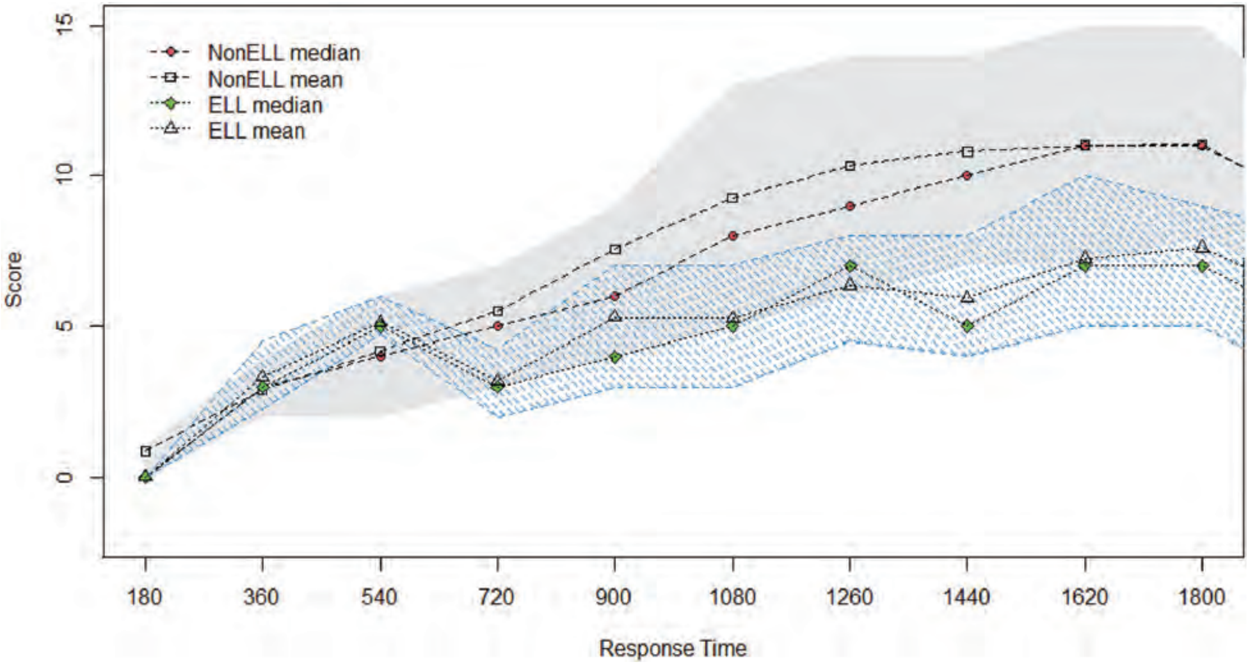

Figure 5 Score comparisons.

We also compared the total scores for each total time subgroup by discretizing the total testing time every 3 minutes. For example, the first subgroup spent less than 3 minutes on the test, the second subgroup spent 3–6 minutes on the test, and so on. The 3-minute time interval was selected to aggregate more test takers in each time subgroup, particularly for the ELL students, so that the comparison was more reliable.

Figure 5 shows that, on average, both ELL and non-ELL students' average scores increased slightly and steadily as their total time increased on the math test (except for the last couple of minutes for non-ELL students). Again, as in Figure 4, the shaded or hatched areas are the interquartile ranges connecting different time subgroups.

Score comparison for each given total testing time subgroup between the English language learner (ELL) and non–ELL students. The gray area is the interquartile range of the non–ELL group, and the blue hatched area is that of the ELL group.

Similar to Figure 5, Figure 6 also shows the score patterns of the ELL and non-ELL groups with the total response time but produced by the nonparametric LOESS method (i.e., the total response time was not discretized but treated as a continuous variable; Cleveland et al., 1992). In Figure 6, the smoothed curves made of points (red diamond for ELL and black dot for non-ELL) are the estimated regression function of scores against response times and its confidence band (made of smaller points). Again, with longer total times, both ELL and non-ELL groups' scores showed slightly and steadily increasing increments, except in the last few minutes. The association between the total scores and the total response times was very low (Spearman correlation coefficient .02 vs. −.02, though significant) for the ELL group versus the non-ELL group. Because of potential speededness, after removing those whose total response times were longer than 29 minutes, the association between scores and response times increased to .25 for the ELL group and .16 for the non-ELL group.

Item-Level Results

As an example, Figure 7 shows the item response time comparison between the ELL and non-ELL groups for Item 19, which can be read as for Figure 4. Other items have similar conditional response time distributions and are not presented.

The gray area is again the interquartile range of the non-ELL group, and the blue hatched area is that of the ELL group. Response time distributions are skewed to the right because the means (triangles for ELL and squares for non-ELL) are larger than the median (diamonds for ELL and circles for non-ELL). The spikes in the figures are caused by very small sample sizes at those score points.

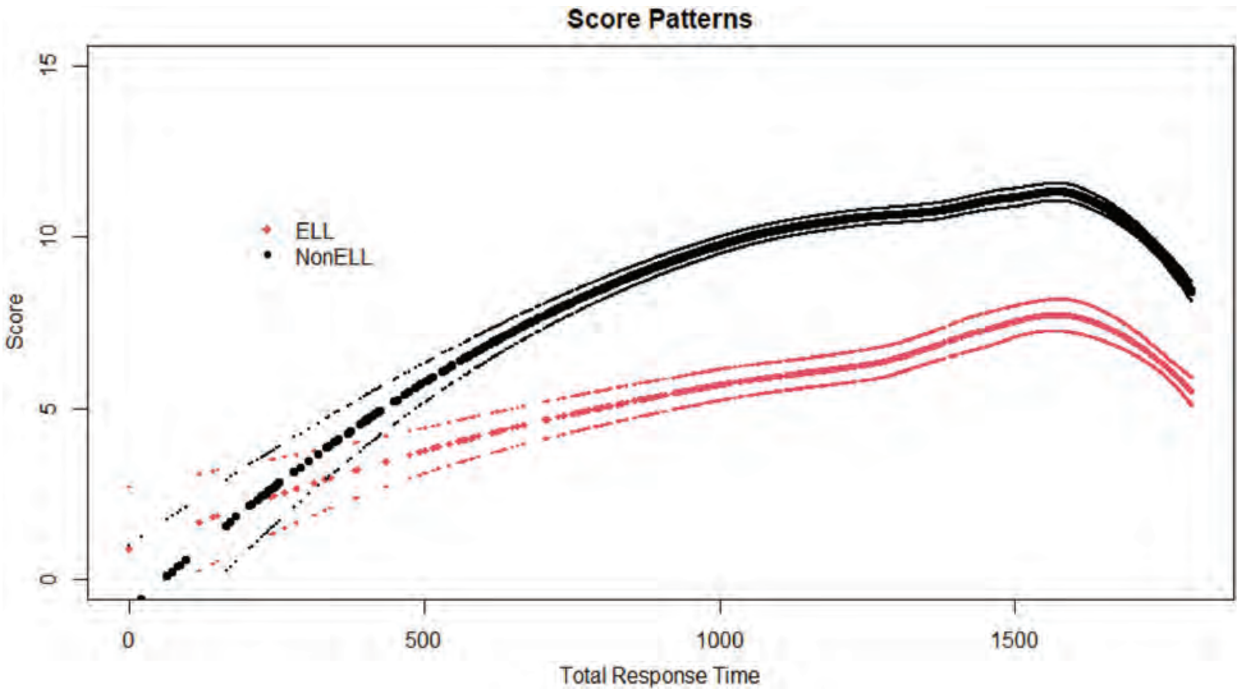

Figure 6 The total score patterns of the English language learner (ELL; red) and non – ELL groups (black) against the total response times, produced by the nonparametric locally weighted scatterplot smoothing method.

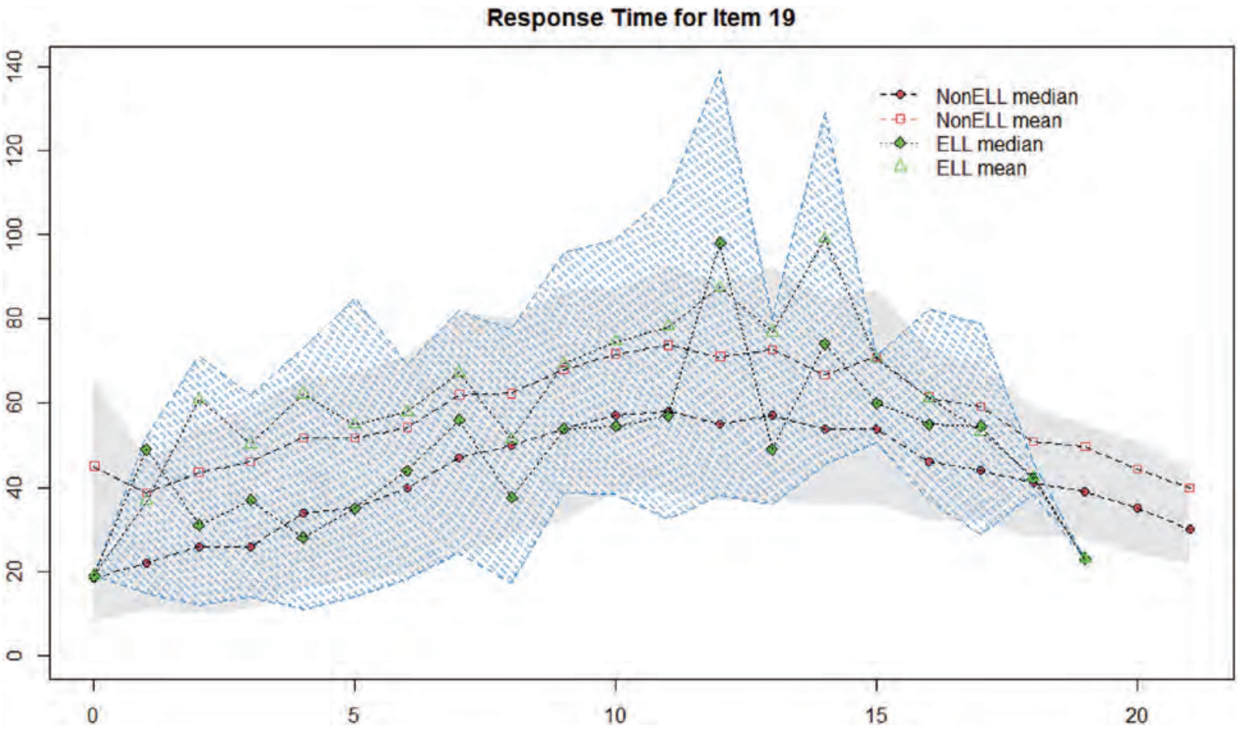

Figure 7 Item response time for Item 19 conditioned on sum scores. The gray area is the interquartile range of the non – English language learner (ELL) group, and the blue hatched area is that of the ELL group.

Table 4 Differential Item Functioning and Differential Response Times (DRT) Between English Language Learners and Non-English Language Learners

	SMD			DRT_S			DRT_T		
Item	Effect size	Test stat.	*p*-Value	Effect size	Test stat.	*p*-Value	Effect size	Test stat.	*p*-Value
1	0.04	2.25	.01	3.52	5.82	.00	6.01	7.94	.00
2	0.00	0.02	.49	8.03	3.40	.00	9.07	2.87	.00
3	0.01	0.48	.31	6.77	1.74	.04	3.91	−0.42	.34
4	−0.01	−0.67	.25	15.22	3.75	.00	31.52	7.59	.00
5	−0.04	−3.20	.00	−1.90	−1.03	.15	−16.96	−9.21	.00
6	0.02	1.40	.08	1.73	1.15	.13	1.37	0.35	.36
7	−0.09	−2.90	.00	11.38	4.66	.00	22.69	8.09	.00
8	0.04	2.75	.00	4.44	2.27	.01	−7.58	−4.26	.00
9	0.01	0.51	.31	3.66	1.83	.03	−7.83	−5.41	.00
10	0.02	0.91	.18	−0.25	−0.79	.21	−3.34	−5.17	.00
11	0.02	1.61	.05	1.22	−0.34	.37	−5.62	−5.82	.00
12	0.01	0.48	.31	1.52	0.77	.22	−2.85	−1.93	.03
13	−0.01	−0.59	.28	7.45	2.44	.01	−9.85	−4.86	.00
14	0.03	2.55	.01	−1.02	−2.47	.01	−10.78	−8.89	.00
15	−0.01	−0.53	.30	−0.27	−1.53	.06	−3.40	−5.65	.00
16	−0.01	−0.54	.30	0.83	−0.74	.23	0.80	−1.92	.03
17	−0.03	−1.82	.03	−0.29	−1.24	.11	−1.74	−3.84	.00
18	0.02	1.03	.15	0.31	−1.56	.06	−7.06	−7.84	.00
19	−0.03	−1.51	.07	4.50	−0.18	.43	−0.72	−4.63	.00

Note. DRT_S uses score matching, and DRT_T uses total time matching. SMD = standardized mean difference.

Using Equations 1–4, we obtained DRT results for the 19 items on Form 2. Table 4 shows the DIF and DRT results between the ELL and non-ELL groups for each of the 19 items. The regular item DIF effect size is shown in the SMD columns of the table (Zwick & Thayer, 1996), with its test statistic and the corresponding *p*-value. Both the DIF effect size and the significance test statistics are used in practice in DIF analyses. As a rule of thumb, the value of .10 is often used as a meaningful effect size to flag items that may have DIF (Dorans & Kulick, 1986; Kim et al., 2007; Zwick, 2012). Table 4 shows that none of the items' SMD is larger than .10 in absolute value. Therefore, DIF may not be a practical concern for score comparison on this form.

Table 4 also reports DRTs for the items. The DRT effect sizes are computed using Equation 1. To produce test statistics and their *p*-values, we took the logarithm transformation of the item response time so that the test statistics more quickly converge to a normal distribution. The results show that DRT_S effect sizes (in the fourth column) with statistical significance are mostly positive and up to 15 seconds. This may indicate that, given similar scores, ELL students spent a longer time on the items than non-ELL students did. Unlike DRT_S, the DRT_T effect sizes (in the seventh column) are mostly positive (except Item 5) on the first half of the items, and then are negative on the second half of the items. That is, given similar total response times, the ELL group may spend a longer time on the first half of the items and a shorter time on the second half. In addition, most DRT_T are statistically significant. Figure 8 visualizes the effect sizes in Table 4.

Summary and Discussion

ELL students' mathematics performance is often affected by their English language skills (Ercikan et al., 2015), particularly on mathematics questions loaded with heavy text. Using timing data collected from a mathematics assessment as an example, we showed that the DRT method, an extension of the DIF methodology, could be used to describe how ELL students progressed through the test in terms of response times. The proposed DRTs capture timing differences at the item level between the ELL and non-ELL groups.

Overall, none of the items can be classified as DIF in practice. However, the ELL group scored much lower than the non-ELL group. The ELL group also spent a slightly longer time on the mathematics form than the non-ELL group (the difference of the total response time is 120 seconds on average). The total response time distributions for both groups were similar: Most students spent fewer than 30 minutes on the test form, but a higher percentage (36%) of ELL students than

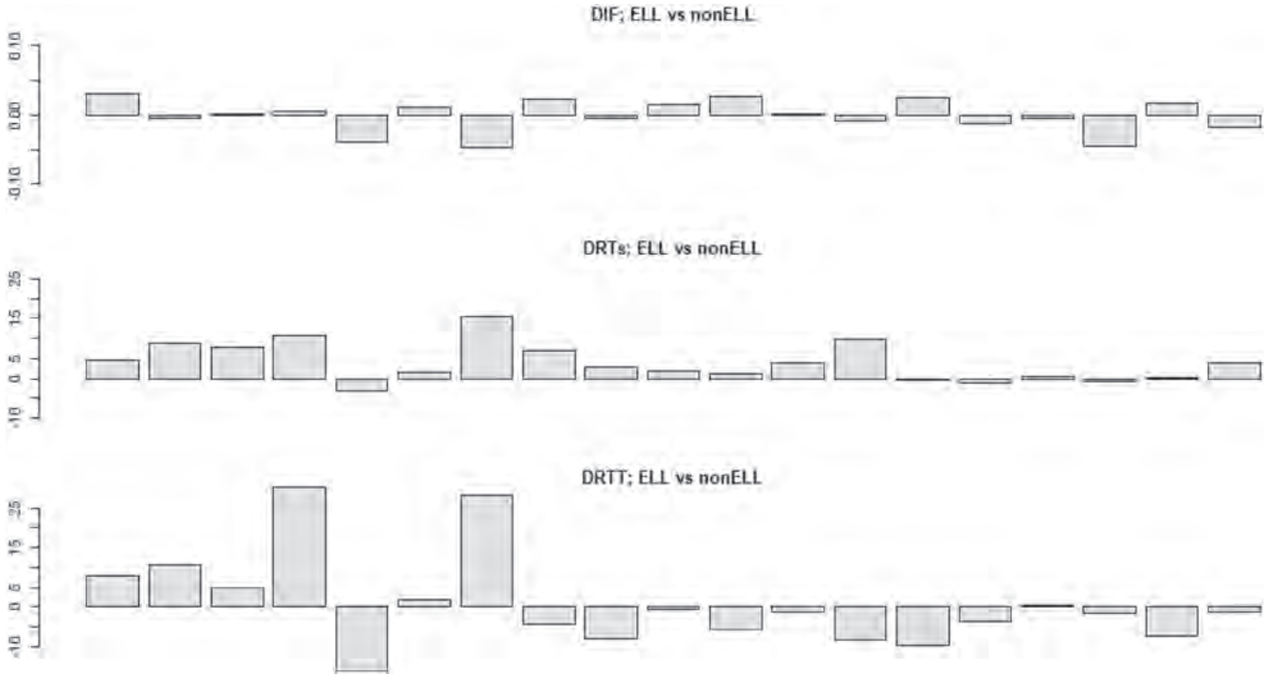

Figure 8 The effect sizes of differential item functioning (DIF; top), DRT$_S$ (middle), and DRT$_T$ (bottom) for English language learner (ELL) and non-ELL groups.

non-ELL students spent about 30 minutes, and ELL students had higher numbers of not-reach items. Hence there may be test disengagement and speededness issues for the ELL group. In addition, both the ELL and non-ELL students who spent a longer time on the form tended to score higher when not under time pressure.

At the individual item level, when matched on their total scores on the form, the ELL students spent a significantly longer time than the non-ELL students on most of the items; the test seemed more time demanding in the English language for ELLs than for non-ELLs. When matched on their total response times on the test form, ELLs tended to spend a longer time on the first half of the items and a shorter time on the second half of the items. The DRT$_T$ results show that the ELLs who used almost all the allotted time might have found the forms speeded and therefore rushed through the test.

The large DRT$_S$ may indicate that ELL students may need a longer time on the mathematics items than non-ELL students; that is, extending testing time for ELL students might help improve their mathematics performance. However, given that most ELL students did not use up all the allotted testing time and given their low scores, extending testing time may not be the most effective way to improve ELL students' performance on the mathematics assessment, in agreement with previous findings (Ercikan et al., 2015).

It is important to develop mathematics items to engage both ELL and non-ELL students on large-scale assessments so that participation of ELLs will be meaningful and valid on these assessments. As Abedi (2014) discussed, using computer technology to provide language-based accommodations may be effective for ELL students. Tools, such as audio, dictionary, or translation, embedded in the computer-based test delivery system may help ELL students understand item contexts better so that they can apply their mathematics skills on the assessment. Further studies using process data collected by computer may help evaluate the effects of these tools.

Our study has a few limitations. First, because the studied form was short, containing 19 items, the total score may not be reliable enough to match students' mathematics skills. As in a DIF analysis, longer tests with larger samples, especially for the focal group, are preferred to study DRT; some adjustment of the matching variable may be helpful (Guo & Dorans, 2019a, 2019b, 2020). Second, even though a second form of the studied mathematics assessment was analyzed, and the findings were similar, more analysis is needed on more forms and more mathematics assessments to evaluate whether similar results hold. Third, in the operational analysis of the studied mathematics assessment, students' sampling weights were used because of complex sampling methods; those weights were not used in our analysis, which may lead to underestimated standard error of score means and thus smaller p-values in DIF and DRT analyses. In addition, because of the low-stakes nature of this mathematics assessment, the total response time may be contaminated by both test disengagement and test speededness. Future studies may also examine whether process data can disentangle the two issues.

Acknowledgments

The authors are very grateful to John Mazzeo, Rebecca Zwick, and John Donoghue for their review and comments on the manuscript and to Neil Dorans and Matt Johnson for helpful discussions on relevant research topics. They also thank Ru Lu and Amanda Papa for providing data and test information.

References

Abedi, J. (2014). Accommodations in the assessment English language learners. In A.J. Kunnan. (ed.), *The companion to language assessment: Volume III, evaluation, methodology, and interdisciplinary themes*. 1115–1129. Wiley

Abedi, J., & Herman, J. (2010). Assessing English language learners' opportunity to learn mathematics: Issues and limitations. *Teachers College Record*, *112*, 723–746.

Cleveland, S., Grosse, E., & Shyu, M. (1992). Local regression models. In J. M. Chambers & T. J. Hastie (Eds.), *Statistical models in S* (pp. 309–376). Wadsworth & Brooks/Cole.

Dorans, N., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*(4), 355–368. https://doi.org/10.1111/j.1745-3984.1986.tb00255.x

Dorans, N., & Schmitt, A. (1991). *Constructed response and differential item functioning: A pragmatic approach* (Research Report No. RR-91-47). ETS.

Dorans, N., Schmitt, A., & Bleistein, C. (1988). *The standardization approach to assessing differential speededness* (Research Report No. RR-88-31). ETS. https://doi.org/10.1002/j.2330-8516.1988.tb00287.x

Drasgow, F. (1986). Polychoric and polyserial correlations. In N. Johnson & S. Kotz (Eds.), *Encyclopedia of statistical sciences* (pp. 68–73). John Wiley.

Ercikan, K., Chen, M., Lyons-Thomas, J., Goodrich, S., & Sandilands, D. (2015). Reading proficiency and comparability of mathematics and science scores for students from English and non-English backgrounds: An international perspective. *International Journal of Testing*, *15*(2), 153–175. https://doi.org/10.1080/15305058.2014.957382

Ercikan, K., Guo, H., & He, Q. (2020). Use of response process data to inform group comparisons and fairness research. *Educational Assessment*, *25*(3), 179–197. https://doi.org/10.1080/10627197.2020.1804353

Guo, H. & Dorans, N. (2019a). *Observed scores as matching variables in differential item functioning under the one-and two-parameter logistic models: Population results* (Research Report No. RR-19-06). ETS. https://doi.org/10.1002/ets2.12243

Guo, H. & Dorans, N. (2019b). *A note on using weighted sum scores in the P-DIF statistics* (Research Report No. RR-19-32). ETS. https://doi.org/10.1002/ets2.12268

Guo, H., & Dorans, N. (2020). Using weighted sum scores to close the gap between DIF practice and theory. *Journal of Educational Measurement*, *57*(4), 484–510. https://doi.org/10.1111/jedm.12258

Guo, H., & Ercikan, K. (2021). Differential rapid responding across language and cultural groups. *Educational Research and Evaluation*, *26*, 302–327. https://doi.org/10.1080/13803611.2021.1963941

Kim, S., Cohen, A., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement*, *44*, 93–116. https://doi.org/10.1111/j.1745-3984.2007.00029.x

National Center for Education Statistics. (2019). NAEP accommodations increase inclusiveness. https://nces.ed.gov/nationsreportcard/about/accom_table.aspx

Schmitt, A., Dorans, N., Crone, C., & Maneckshana, B. (1991). *Differential speededness and item omit patterns on the SAT* (Research Report No. RR-91-50). ETS. https://doi.org/10.1002/j.2333-8504.1991.tb01417.x

Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (Research Report No. RR-12-08). ETS. https://doi.org/10.1002/j.23338504.2012.tb02290.x

Zwick, R., & Thayer, D. (1996). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics*, *21*, 187–201. https://doi.org/10.3102/10769986021003187

Suggested citation:

Guo, H., & Ercikan, K. (2021). *Comparing test-taking behaviors of English language learners (ELLs) to non-ELL students: Use of response time in measurement comparability research* (Research Report No. RR-21-25). ETS. https://doi.org/10.1002/ets2.12340

Action Editor: John Mazzeo

Reviewers: Rebecca Zwick and John Donoghue

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/