EDUCATIONAL ASSESSMENT
https://doi.org/10.1080/10627197.2020.1804353

Routledge
Taylor & Francis Group

Use of Response Process Data to Inform Group Comparisons and Fairness Research

Kadriye Ercikan, Hongwen Guo, and Qiwei He

ETS, Princeton, New Jersey, USA

ABSTRACT

Comparing group is one of the key uses of large-scale assessment results, which are used to gain insights to inform policy and practice and to examine the comparability of scores and score meaning. Such comparisons typically focus on examinees' final answers and responses to test questions, ignoring response process differences groups may engage in. This paper discusses and demonstrates the use of response process data in enhancing group comparison and fairness research methodologies. We propose two statistical approaches for identifying differential response processes which extend the differential item functioning (DIF) detection methods and demonstrate the complementary use of process data in comparing groups in two case studies Our findings demonstrate the use of response process data in gaining insights about students' test-taking behaviors from different populations that go beyond what may be identified using response data only.

Introduction

Central to the interpretation of responses to assessment tasks is the extent to which responses, such as selection of a choice on a multiple-choice test, solutions to problems, or essays, could be used as evidence of levels of proficiency in knowledge, skills, attributes targeted by the assessment. Assessments are typically carefully designed to allow such inferences. However, there is an inferential gap between the targeted theoretical constructs and the examinee responses. Examinees' final answers to assessment tasks, "responses" or "products", may not provide information about what kind of response processes examinees engaged in to reach their responses, it may not be possible to determine whether the examinees engaged with the tasks in the intended ways, and whether different examinees took different solution strategies to produce the same responses. Differences in response processes that are not captured in final solutions and responses may be important indicators of levels and types of proficiency, as well as how students from different backgrounds engage with the assessment. In particular, response processes may provide insights that can help build bridges between the observed examinee products and the targeted constructs and provide supporting inferences about the performance. These insights have heightened importance if intended inferences include comparisons of group performances such as whether students from different cultural, ethnic or language backgrounds are engaging tests in similar ways and whether performances by the compared groups can be interpreted in the same way.

The importance of response processes in assessment has been highlighted for a very long time. We see evidence of this recognition in the first edition of Educational Measurement (Lindquist, 1951), in a chapter by Ryans and Frederiksen on performance assessment. Ryans and Frederiksen (1951) addressed the question of "Should measurement of performance be directed at the process or the product?" highlighting the importance of using both:

CONTACT Kadriye Ercikan kercikan@ets.org ETS, Princeton, New Jersey, USA

> While performance may often be measured either in process or in terms of the product, some situations are more limited and offer less choice. Occasionally the product of performance *is not* measurable apart from the performance in process. Playing a musical instrument is an example of such a case. … In other instances, *both* the performance in process and the product of the performance may be measured by judging an individual's proficiency at a given sort of behavior … it is highly desirable to measure actual operation, or performance in process, in many cases. The final product of performance may appear to be of satisfactory quality, but operational methods or procedures employed may have been unsatisfactory … (pp. 471–472).

In recent years, the focus has been on how response processes can inform measurement based on student responses or products, which are the product-based scores. The importance of using response process data for such uses, hereby referred to as process data, has been highlighted in the last two editions of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999, 2014). The emphasis in the Standards has been on the use of process data in validating meaning of product-based scores. Researchers echoed such use for over two decades (Ercikan, 2006; Ercikan & Pellegrino, 2017; Haertel, 1999; Leighton, 2017; Oranje, Gorin, Jia, & Kerr, 2017; Pellegrino, Chudowsky, & Glaser, 2001). Research on curricular standards and for understanding how people learn shifted focus from validating meaning of product-based scores to an emphasis on the inseparability of process and product as recognized by Ryans and Frederiksen (1951). In particular, centrality of "process" in complex constructs in new curricular standards, such as the Common Core, and the emphasis on 21st century skills, including problem solving, reasoning, collaborative problem solving, collaboration and interactive communication, brought the importance of response processes to the forefront, thereby going beyond complementing product-based measurement. In addition, research on how people learn revealed important potential role for assessments to inform learning (Black & Wiliam, 1998; Pellegrino et al., 2001). Central to this role is information about how students engage with assessment tasks, what processes they follow, what kinds of errors they make in coming up with their solutions, all focusing on response processes rather than final solutions.

Despite these emphases on response processes over several decades, use of process data in assessment research and practice has been limited. In paper-based assessments, information about response processes have been gathered through a methodology called think aloud protocols (Ercikan et al., 2010; Leighton, 2017). These methods engage students in thinking aloud as they engage with assessment tasks. Individual administration of think aloud protocols and labor-intensive analyses of the verbalizations required these types of research to be based on small samples of students. With digital assessments replacing many paper-based assessments, data on large numbers of students' actions and response processes can be captured in computer log files. These data provide opportunities for analyzing response processes that have not been possible in paper-based assessments. Such data include how long examinees spent on different parts of assessments, how they navigated through the assessment, what steps and actions they took in responding to items, and whether and how they utilized tools provided by the assessment, and their gazes and the timing of their gazes on the screen (e.g., Goldhammer, Naumann, & Greiff, 2015; He & von Davier, 2016). Data collected with advanced technology such as eye tracking enable fine-grained observation of respondent engagement at the item level. These details provide important clues to visual attention (Ferreira & Henderson, 2004) as well as insightful data regarding assessment response processes that are not obtained from assessment log files, think aloud protocols or administrator observations of assessment. Eye tracking techniques are extensively used in psychological research in areas such as reading, joint gaze and scene perception (e.g. Clifton et al., 2016; Lai et al., 2013; Liversedge, Schroeder, Hyönä, & Rayner, 2015; Oranje et al., 2017) and gradually applied in large-scale assessments. For instance, Maddox et al. (2018) used eye tracking techniques to make detailed observations of item response processes in the digital problem-solving tasks in the International Assessment of Adult Competencies (PIAAC). The lab-based study also recorded physiological responses using measures of pupil diameter and electrodermal activity. The eye tracking observations helped to fill an "explanatory gap" by providing data on variation in item response processes.

Important to note is that, even in digital assessments, cognitive response processes themselves are not observable. What are captured in the think-aloud protocols as well as in log files need to be considered as "traces of processes" rather than processes themselves. Interpretation of these traces, guided by cognition models, can provide important insights about the assessment tasks and measurement in two related ways. One is in the context of assessments that produces the product-based scores, and the other is in the context of assessments that derives scores from students' response processes and sequences of actions. In the context of assessments on final products, response process data can enrich the assessment by providing insights about whether students engage with tasks in the intended ways and thereby contribute to improving assessment design and validating score meaning. Such use is central to assessments of complex constructs that may involve cognitive processes explicitly identified in the targeted construct. These include assessments of reasoning, critical thinking, computational thinking, etc. In the context of assessments where sequences of actions are used in the creation of scores, students' actions are treated as evidence of the targeted construct. This involves assessments of constructs where sequences of actions are central to assessment such as in the case of interpersonal communication, problem solving, and cross-cultural competence.

The importance of response processes is heightened when assessments are used in comparing groups. One goal of comparing groups is to gain insights to inform policy and practice, and the other one is for examining the comparability of scores and score meaning for the comparison groups (Ercikan, Roth & Asil, 2015). When such comparisons are restricted to student products and ignore response processes, significant gaps may arise in understanding whether students from different groups engaged in assessment tasks in similar ways, whether students adopted different solution strategies, and what kinds of response processes examinees engaged in coming up with their solutions. Response process data can provide insights on how students from different groups engage with assessment tasks and can support comparisons of performance levels across groups as well as comparability of measurement across groups.

In this paper, our goal is to discuss and illustrate the use of response process data in enhancing methodologies used in comparing groups. We demonstrate how response processes may reveal important information about differences that may not be captured by the final responses. We argue for the use of response process data in addition to final responses to test questions in comparing groups and for examining measurement comparability. We propose methodologies and demonstrate the use of process data in comparing groups in two case examples.

Study 1: use of timing data in comparing ELL and non-ELL performance on a mathematics assessment

There are consistent patterns of differences in performance levels of English Language Learner (ELL) students and non-ELL students on large-scale assessments such as state assessments, national and international assessments (Abedi & Herman, 2010; Ercikan et al., 2015; Ercikan & Elliott, 2016; Willner, Rivera, & Acosta, 2007). These differences may be due to a complex set of factors such as differences in opportunities to learn, experience with assessment practices, and the language demands in the assessment (Menken, 2008). In this study, we used item responses and item response time to examine differences between the ELL and non-ELL students on an eighth-grade mathematics assessment. The main purpose of the study was to gain insights on response processes of ELL students as compared to non-ELL students on an assessment that was not intended to assess ELL students' language proficiency in English. Such differences in response processes may have implications on score meaning, validity, and comparability for the two comparison groups.

Differences in response times can be important indicators of differences in test-taking effort, motivation, difficulty level of the items, and test speededness (Guo et al., 2016; Rios, Guo, Mao, & Liu, 2017; Van der Linden, 2011; Wise, 2017). In addition to variations in response times due to an interaction of this complex set of factors, response times are expected to vary by proficiency level and as a result of individual differences in how students pace themselves through responding to the

assessment (Wise & Kong, 2005). We propose two statistical approaches for identifying differential response processes which are extensions of differential item functioning (DIF) detection methods (Dorans & Kulick, 1986; Zwick, et al., 1996). Using this methodology, we investigate whether there are response time differences when students are matched on (1) their mathematics scores or (2) their total response time on the test. Identification of differential response time matched on mathematics score can highlight items that may require additional time to read and understand or compose responses for students from different populations who are expected to perform at the same level. Differential response time for populations matched on total response time, on the other hand, may point to items which need differential time for groups who complete the test at the same pace. Both differences can be considered as aspects of differential response processes for the considered populations. In identifying sources of such differential response processes, it is important to note that spending more time on an item may be due to multiple factors, including additional cognitive demands to understand language, item format, and content and context presented in the item.

Data

The data in Study 1 were collected from 27,622 students who responded to the items in a large-scale mathematics assessment. There were 3,149 ELL students and 25,653 non-ELL students in the sample. The mathematics assessment was part of a large-scale survey in North America that was intended to measure student groups' proficiency in various subjects; this assessment was chosen for our study because it had large sample sizes and represented the large population of the eighth graders. The mathematics assessment had 15 items that consist of different item types (such as the multiple-choice items with single or multiple sections, and the constructed response, i.e. CR, items). The total raw score for the assessment ranged from 0 to 26. Items in this assessment was considered to be innovative compared to the traditional multiple-choice items, and they included multiple item formats with graphs, tables, and figures. In addition, the multiple-choice items had varying number of options. An example item is presented in Figure 1 that uses figure and drag-and-drop to facilitate students' responses, and it has fifteen possible combinations for students to choose from.

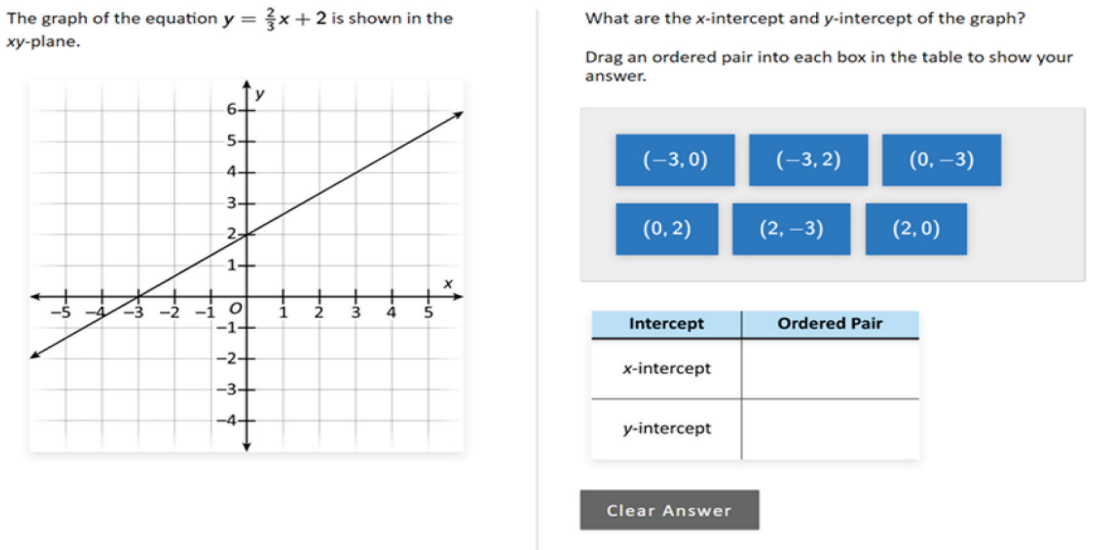

Figure 1. An example item on the mathematics assessment.

Table 1. Item and test information of the assessment in study 1.

Item	Max Score	Average percent score	Polyserial correlation	Median RT	Content
1	1	0.65	0.72	29	Translate a percent to a fraction in context
2	1	0.94	0.45	56	Complete a circle graph to represent data
3	1	0.47	0.58	72	Find the product of two two-digit decimals
4	2	0.61	0.68	57	Determine the x/y-intercept
5	2	0.73	0.71	80	Compare measurements using unit conversion
6	2	0.22	0.7	84	Extend a numerical pattern
7	1	0.12	0.33	77	Determine the diameter of a circle
8	1	0.37	0.43	51	Identify a verbal description of a rotation
9	1	0.67	0.65	61	Create a proportion
10	2	0.27	0.64	67	Identify characteristics of lines
11	2	0.15	0.7	134	Make and explain a conclusion about two linear
12	2	0.12	0.63	63	Identify figures
13	4	0.24	0.72	175	Evaluate a circle graph and bar graph to Cluster
14	2	0.28	0.55	79	Match box plots to given stem-and-leaf plot
15	2	0.12	0.77	86	Create an expression for the area

Table 1 summarizes item statistics and presents brief descriptions for the 15 items on the assessment. The maximum score on each item ranged between 1 and 4. The percent of maximum score (average item score divided by maximum score), an indicator of difficulty level, ranged from .12 to .94, demonstrating that difficulty level of the items varied widely, and the assessment included some very difficult items as well as very easy ones. The item discrimination coefficients (item polyserials; Drasgow, 1986) ranged from .33 to .77, indicating relatively strong relationship between individual items and overall performance on the assessment. The item response times (RT), defined as the total time a student spent on the item, varied across items, and the median testing times[1] ranged from 29 to 177 seconds.

Table 2 presents summary statistics for scores, reliability and response times for ELL and non-ELL students. Results indicate a statistically significant and large, approximately one standard deviation, performance difference between ELL and non-ELL students, with ELL students performing much lower. The reliability estimates demonstrate a higher measurement accuracy (.79 compared to .61) for non-ELL students; the lower reliability of ELL students is most likely due to a restricted score range of ELL students as shown in the score standard deviation There was also a difference in the total testing time for ELL and non-ELL groups, with ELL students spending significantly greater time in responding to the test. For further insights about response time, Figure 2 shows the histograms of the total response time (TRT) for the ELL and non-ELL students. As can be seen, the total time distributions have similar patterns for the two groups: they are bi-modal distributions with a high mode at the end of the assessment. The two modes (in seconds) are 1075 and 1751 for the ELL group, and they are 1110 and 1748 for the non-ELL group. A mixture with two normal distributions was fit to the TRT distribution, which shows that about 22% ELL students and 18% non-ELL students belong to the right high-modal-TRT cluster who spent 1690 to 1800 seconds on the test block. That is, most of the students (approximately 80% in both ELL and non-ELL groups) belong to the low-modal-TRT cluster and did not use all the allotted time on the math block, and the ELL group spent slightly longer time than its counterpart on average.

Table 2. Comparison scores, reliability and response time for ELL and non-ELL.

	Non-ELL (N=23741)	ELL (N=1237)
Total score mean (SD)	9.20 (4.74)	5.45(3.07)
Reliability (Cronbach's alpha)	.79	.61
Total time median (IQR)	1376 (535)	1408(564)

Note: Response time (RT) was recorded in seconds and ranged from 0 to 1800 seconds. Because of the skewness of the RT distribution, we considered median and IQR.

[1]Median of response time is used because of the highly skewed nature of the timing distributions.

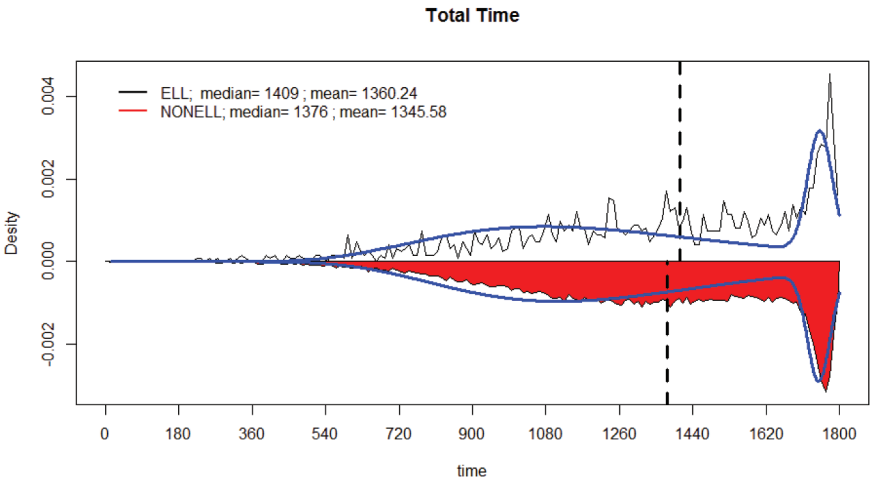

Figure 2. Histograms of the total response time for the ELL group (upper curve) and the non-ELL group (lower curve). Note that the test block had a maximum time of 1800 second (i.e., 30 minutes).

Figure 3 shows the median and mean scores of the ELL and non-ELL groups plotted against time spent on the test in seconds. The gray area with dashed lines in the center in Figure 3 is the inter-quartile range of scores for the non-ELL group, and the blue area with dotted lines in the center is that of the ELL group. As can be seen on the diagram, on average, the non-ELL students' scores (squares in the gray area) increased steadily as their total time increased on the math test except for the last two minutes; the ELL students' scores (triangles in the blue area) increased with some fluctuations. As expected, the non-ELL students outperformed the ELL students on average for any given total response time.

In an effort to examine the differences between ELL and non-ELL students in different TRT clusters, we compared their performance levels and relationship between response time and performance. Table 3 shows the score summary statistics for students whose TRTs belong to one of the two clusters (i.e., the low-modal-TRT cluster and the high-modal-TRT cluster), and the correlation of scores with TRT, respectively. The score means in the high-modal-TRT clusters are significantly higher than those in the low-modal-TRT clusters for both ELL and Non-ELL, indicating students who

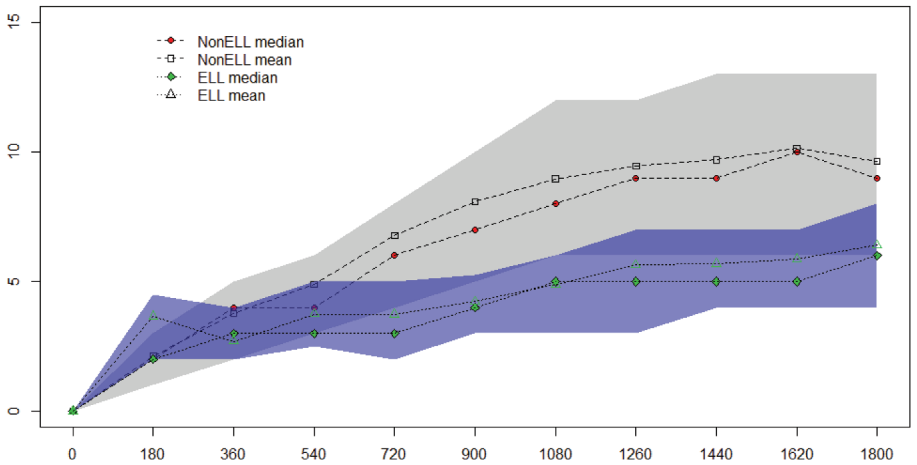

Figure 3. Median and mean score comparison for each given total response time (TRT) group between the ELL and non-ELL groups. The x-axis is TRT in seconds, and the y-axis is score.

Table 3. Scores for students in different TRT clusters.

			Score Mean (SD)	Corr (TRT, Score)
ELL	All	1237	5.45 (3.07)	0.29*
	Low-modal-TRT	936	5.17 (2.94)	0.29*
	High-modal-TRT	301	6.35 (3.28)	0.04
non-ELL	All	23741	9.20 (4.74)	0.20*
	Low-modal-TRT	18712	9.06 (4.78)	0.23*
	High-modal-TRT	5029	9.75 (4.53)	−0.18*

Note. The asterisk indicates the statistical significance with p-value < .05.

used almost all the allotted testing time scored higher than those students who did not use all the testing time. The relationship between scores and response time, however, varied for the low-modal- and high-modal-TRT clusters. The correlation between test scores and TRT is positive for the low-modal cluster for both ELL and non-ELL students, but it is zero or negative for the high-modal cluster. This low or negative correlation in the high-modal cluster is partly due to the narrow time window (from 1690 to 1800 seconds), which led to very low variance in TRT resulting in low correlations. Another contributing factor for the low or negative correlation in the high-modal-TRT cluster could be due to inclusion of low-performing students who were spending all the allotted time to solve items that may seem too difficult, in addition to high performers who had extra time to review items. On the other hand, among students in the low-modal-TRT cluster, those who spent more time on the test tended to perform better, with a moderate correlation of .29 for ELL and .23 for non-ELL between scores and TRT. The existence of the bi-modal TRT distributions might be an indication of low test engagement for some students because of the low-stakes nature of the assessment.

Differential response time

For identifying differential response time (DRT) for the ELL and non-ELL groups, we used an extension of Differential Item Functioning (DIF). DIF analysis is a standard fairness practice in the testing industry since the 1980s (Dorans, 2013; Ercikan, 1998; Ercikan & Lyons-Thomas, 2013; Holland & Thayer, 1988; Holland & Wainer, 1993; Zieky, 1993, 2011; Zwick, 2012) and is used to identify differential response patterns for groups of students matched on overall performance. A DIF analysis compares whether a test item functions differently for test takers in different groups (focal vs. reference) with the same expected performance levels. The DIF statistic we used is an extension of the generalized version of the standardized proportion difference (STD *P*-DIF, Dorans & Kulick, 1986) – the standardized mean difference (SMD, Zwick & Thayer, 1996). SMD tests whether the average item score is different between the two groups for members of the same ability.

We focused on two statistics that can be used as indicators of differential response processes. Differential Response Time (DRT), conditional on total score (DRT.SS) is defined as the item response time difference between the ELL and non-ELL groups after matching on their total test scores. DRT, conditional on total response time (DRT.TT), on the other hand, is defined as the item response time difference between the ELL and non-ELL groups after matching on their total response time on the test. We obtained DRT.SS and DRT.TT results for the 15 items in the assessment.

Table 4 shows the DIF and DRT results between the ELL and non-ELL groups for each of the 15 items. The regular item DIF effect size identified by SMD (Zwick & Thayer, 1996) is shown in the second column of the table (the asterisk next to the effect size indicates statistical significance at p-value < .05). The value of 0.10 is considered a meaningful effect size used in the operational DIF analysis (Dorans & Kulick, 1986; Zwick, 2010). None of the items' DIF effect sizes is larger than 0.10 in absolute value. Therefore, DIF may not be a concern for the assessment. However, DRT.SS effect sizes (in the third column) are all positive and are between one second and ten seconds on average. Most of the item RT differences are statistically significant (note that the significance test for DRT is different from that for DIF because of the continuous nature of RT. In addition, logarithm transformation of item RT was taken for significance tests). DRT.SS results indicate that ELL students spent slightly longer time on average

Table 4. DIF and DRT effect sizes between ELL and non-ELL.

Item	DIF	DRT.SS	DRT.TT
1	−0.08*	9.48*	13.62*
2	−0.01	8.18*	14.08*
3	0.04*	4.42*	8.31*
4	0.03	8.64*	16.83*
5	−0.01	1.24	3.32*
6	−0.03*	10.33*	14.31*
7	0.04*	9.63*	0.73
8	0.04*	2.85	−0.23*
9	−0.05*	4.37*	1.74
10	−0.02	3.06	−4.72*
11	−0.02*	5.54*	−11.91*
12	0.00	1.18	−10.09*
13	0.00	0.57	−29.78*
14	0.08	4.92	−8.85*
15	0.00	7.33	−10.60*

Note 1. The asterisk indicates the statistical significance with *p*-value < .05.

Note 2. DRT.SS and DRT.TT are the differential item response time after matching on total score and total response time, respectively.

than their counter-part non-ELL students who had similar total scores. In addition, these differences on half of the items are statistically significant. Unlike DRT.SS, the DRT.TT effect sizes are positive in the first half of the fifteen items, and they are negative in the second half of these items. Furthermore, most of the differences are statistically significant. DRT.TT results indicate that given similar total response time spent on the math block, ELL students spent relatively longer time on the first half of the items and shorter time on the second half of the items, compared to their non-ELL counterparts, pointing to possible greater degree of speededness for ELL students. These patterns are visualized in Figure 4 which displays DIF, DRT.SS, and DRT.TT effect sizes.

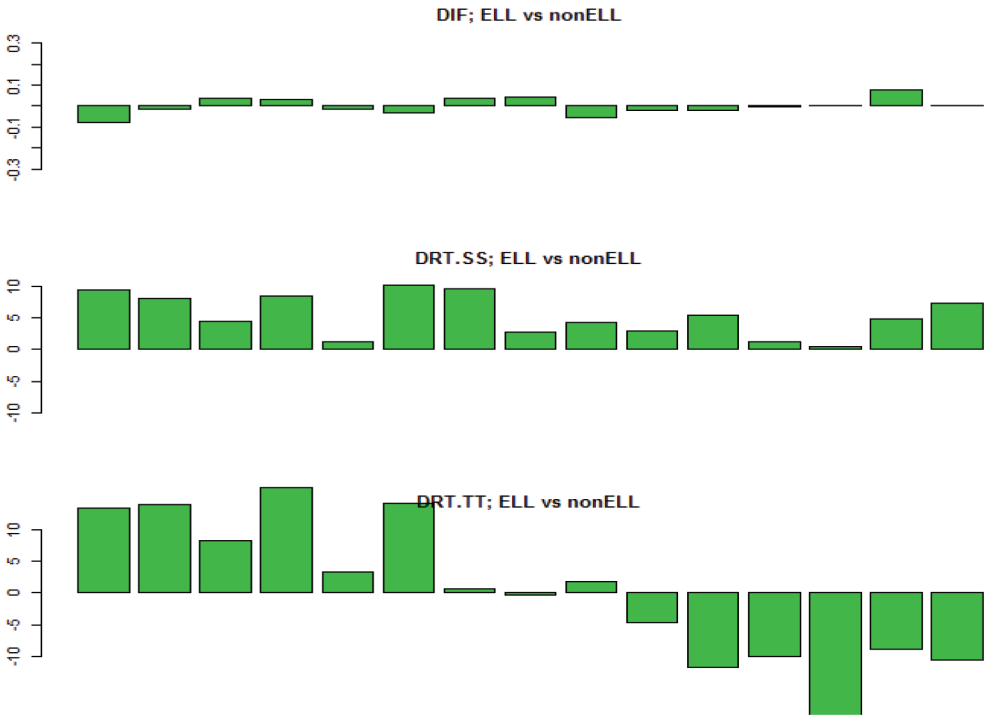

Figure 4. The effect sizes of DIF (upper panel), DRT.SS (middle panel), and DRT.TT (lower panel) between ELL and non-ELL groups.

In view of the bi-modal distribution of the total response time, we investigated the differential response patterns between ELL and non-ELL groups for students in the high-modal cluster who spent nearly all the allotted time (DRTs for the low-modal-TRT clusters' comparisons were similar to the total group comparison, mainly due to the fact that the low-modal-TRT clusters were about 80% of the total groups). For the high-modal-TRT clusters' comparisons, the negative DRT.TT and DRT.SS in the second half of the items in Figure 5 indicate that there may be more speededness for the ELL students than the non-ELL students.

Summary of Study 1

This study shows, overall, the ELL group scored lower than the non-ELL group, and they also spent slightly longer time on the mathematics assessment than the non-ELL group. DIF analyses on item responses (that is, their answers to test questions) indicate that items functioned similarly for the ELL and non-ELL groups. Based on just the DIF findings we may conclude that there is no evidence to suggest that there are threats to score comparability across the two groups. However, when we compared the item response times, one aspect of response processes, we observed differences in response time for the two groups. DIF.SS showed that ELL students took slightly a longer time on most of the items than those with similar ability in the non-ELL group. DIF.TT showed that ELL students took a longer time on the first half of the items and shorter time on the second half of the items, compared to non-ELL students who had similar total testing times.

The findings from Study 1 illustrate the use of timing data in gaining insights about students' test engagement from different populations that go beyond what may be identified by using response data only. In particular, the findings show that ELL students who are at the same mathematics ability/performance level as non-ELL students tend to spend longer time in responding to test questions. This is not a surprising finding, given the lower language proficiency levels for ELLs in English and the expected language demands on all assessments, even on a mathematics assessment, on which the majority of the items were word questions that required English reading ability. However, empirical evidence that demonstrates the degree of difference in response times for ELL at all performance levels highlights the potential impact of such difference on overall performance levels and score

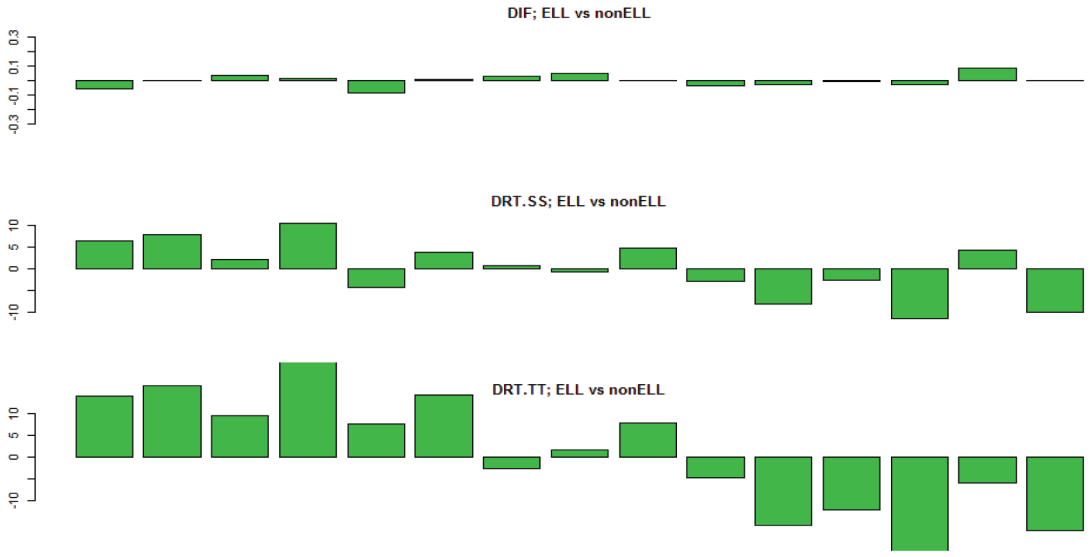

Figure 5. The effect sizes of DIF, DRT.SS, DRT.TT for students in the high-modal-TRT cluster of the TRT distribution between ELL and non-ELL groups.

comparability. ELL students who need more testing time to demonstrate their knowledge, skills and competencies might be disadvantaged if there were speededness in the assessment, particularly for those who used almost all allotted testing time. On the other hand, large positive DRT. TT in the first half of the test block and large negative DRT.TT in the second half show that ELL students distributed relatively longer time on the first half items and shorter time on the second half items than the non-ELL students, after matching on the total response time on the item block. Shorter time distributed on the second half of the test from the ELL group might be traces of low test engagement possibly due to test anxiety and frustration.

For this Math assessment, the traditional DIF analysis did not flag any items as potentially biased for the ELL and non-ELL groups. However, based on timing data, we found differences in DRTs. This case study demonstrates that DRT.SS and DRT.TT may provide evidence regarding the comparability of test responding time distributions for the comparison groups that may not be captured by DIF analysis alone.

There are limitations in this study. Because the studied assessment was short, containing only 15 items, the relatively low reliability (Cronbach's alpha = .79 for the non-ELL group, and .61 for the ELL group) and the small sample of the ELL group may have affected statistical inferences of DIF and DRT analyses. As in DIF analyses, longer tests with larger samples, especially for the focal group, are preferred in examining DRTs. In addition, because the data were collected from operational programs, instead of an experimental design, item difficulty, position effect, and speededness may be entangled for both ELL and non-ELL students, and it is unclear whether extending testing time for ELL students will improve their performance, given the low-stakes nature of the assessment and the fact that most students did not use all the allotted time. Further studies need to address how effective technology-enhanced tools, such as embedded dictionaries and audio translations, may help improve ELL students' performance on mathematics skills.

Study 2: Use of timing data and action numbers in comparing culture group performance on a science assessment

In the second case study, we investigated response process differences among cultural and language groups on the Progamme for International Student Assessment (PISA). PISA is a triennial international survey which aims to evaluate education systems worldwide by assessing the skills and knowledge of 15-year-old students. We explore response process differences between different cultural and language groups to gain insights on performance differences and score comparability.

Data

In the 2015 PISA assessments, Science was the main assessment domain, which assessed students' skills and knowledge in science (OECD, 2018). Data released from 2015 PISA Science assessment (OECD, 2018) contained item responses, item response times, and numbers of actions students took to answer an item. Items were a mixture of multiple choice (MC) items and constructed response (CR) items. For the CR items, students typed their responses on computers through keyboards. Item response time was recorded in seconds, and the number of actions was the count of keystrokes the computer recorded when a student responded to an item. In addition, students' proficiency scores, estimated by the plausible values (PVs, Mislevy & Sheehan, 1987), were available for each student. In 2015, Science assessment contained 12 item blocks/clusters consisting of 184 items, and combinations of two item blocks/clusters were administrated to students in a testing window of one hour. We analyzed the six new item blocks for four cultural groups that had relatively larger sample sizes and different languages. These groups were students from Canada (CAN), China (CHN), Korea (KOR), and USA. Table 5 shows the total sample size, ranking in performance on the science assessment, and language of each country group. Note that Canada, China, and Korea are among the top performing countries, and USA is in an average performing country in Science.

Table 5. The country group information on Science.

Groups	Sample Size	Science Ranking	Language
Canada	15,389	7	English
China	9,805	10	Chinese
Korea	5,581	11	Korean
USA	5,677	25	English

Note. Even though PISA was administered in both English and French in Canada, we only used the data from the English-speaking sample. Sample sizes were obtained after data cleaning.

Methods

The two DRT methods described in Case study 1 were used to analyze item response time differences between the country groups on the science item blocks. In addition, differential item action (DIA) was defined in the similar fashion to DRT to investigate whether students in the focal group had larger or smaller number of actions than those in the reference group after matching on their plausible values.

Results

The upper panel of Table 6 and Figure 6 shows the overall summary of results across the six new science blocks for each studied country. Table 6 summarizes sample size for each country, average performance level as indicated by plausible values (PV), median total response time (TRT), mean (SD) number of actions (Action) and correlations between PV and TRT, Action and TRT and PV and Action. As summarized in Table 6, in terms of TRT, Korean test takers used shorter average total response time (TRT) than test takers in the other countries; in terms of actions, Chinese, and Korean test takers had smaller average numbers of actions than test takers in the English language countries (Canada and USA). The correlation coefficients between PV, TRT, and Actions were different as well among different country groups. Correlations among PV, TRT, and Actions were low for Chinese test takers, and correlation between PV and TRT was low for US test takers. Figure 6 shows the distributions and correlations of the three variables (PV, RT, Action) for each country group across the new blocks, which conveys similar messages to Table 6.

Due to similarity of results of the analyses for the new science blocks and the goal of the case study being to demonstrate how response process data may inform group comparisons, we only present results obtained from one new block (17 items) when it was in the first position of the two-block combination. In addition, in the following analysis, the Canadian group was used as the reference group because it had the largest sample size. The lower panel of Table 6 shows the summary statistics of this new block for the studied countries. As in the upper panel of Table 6, the three variables (PV, TRT, and Action) and their relationships vary among country groups. For simplicity, we use the country name to denote the test taker group from that country.

Table 6. Summary statistics of response and process data for the four countries.

Country	Sample Size	Mean PV (SD)	Median TRT (IQR)	Mean Action(SD)	Correlation (PV, TRT)	Correlation (TRT, Action)	Correlation (PV, Action)
On the six new Science blocks							
Canada	15,389	519(92)	1107(373)	739(537)	0.36	0.60	0.52
China	9,805	527(100)	1200(410)	156(223)	0.20	0.17	0.08
Korea	5,581	515(95)	872(345)	256(216)	0.45	0.46	0.34
USA	5,677	497(97)	1202(400)	717(474)	0.18	0.48	0.48
On the first new Science block							
Canada	1681	516(94)	1176((421)	616(476)	0.32	0.59	0.51
China	1091	528(99)	1336(467)	157(174)	0.15	0.16	0.07
Korea	617	510(98)	893(350)	227(168)	0.45	0.45	0.53
USA	642	497(99)	1306(557)	699(396)	0.07	0.36	0.45

Note. PV stands for plausible value (i.e., proficiency), and TRT is the total response time in seconds on the studied item block.

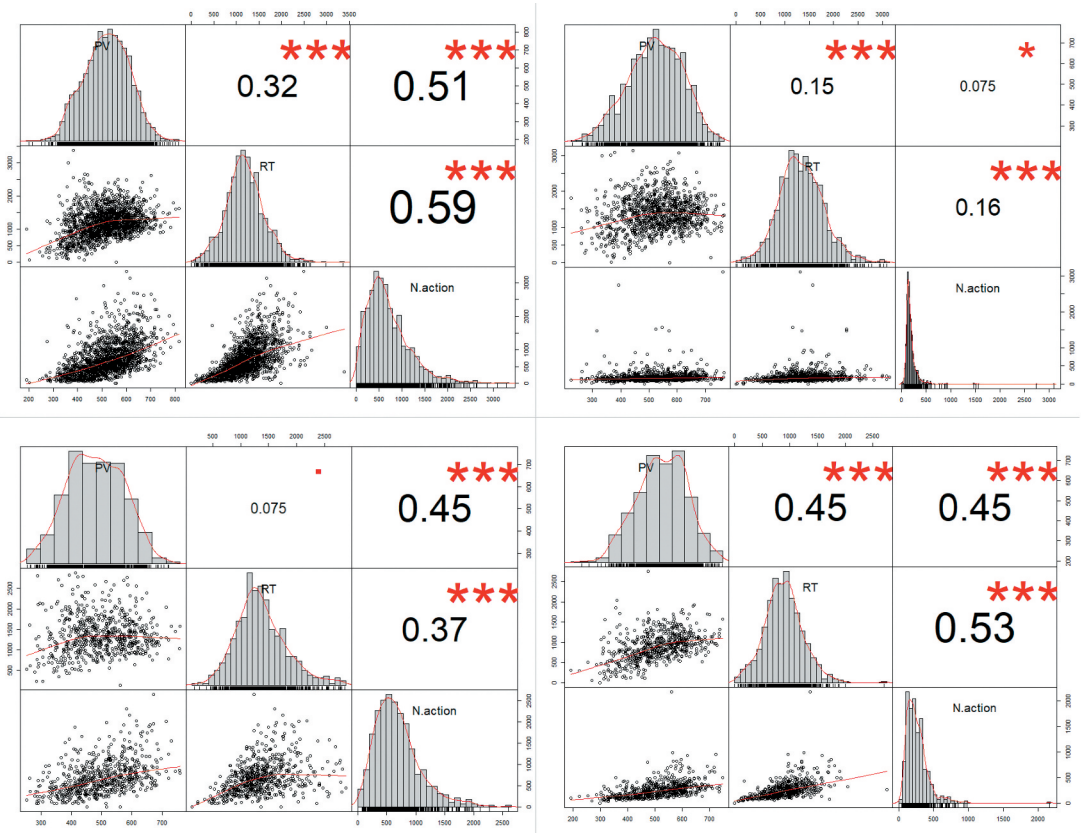

Figure 6. Distributions of and correlations among PV, TRT, and Action for Canada (CAN), China (CHN), Korea (KOR), and USA (from top left, clockwise).

Table 7 shows item summary statistics, DIF, DRT, and DIA results for China (focal group) and Canada (reference group) comparison. The second column shows the average percent score (average item score divided by maximum score; items 4, 8, 16 were polytomously scored as 0, 1, or 2, and the others were dichotomously scored as 0 or 1). The third and fourth columns show the median RT and average numbers of actions of the reference group (Canada) on the item block, respectively. Items 6, 8, 10, and 16, which are the only CR items in this block, had longer median item RT and larger numbers of actions than the other items. The fifth column is the DIF effect size (Zwick, 1990). Items 1, 3, 4, 11, and 14 were flagged as DIF items between the China and Canada groups; these flagged items also had country-specific item parameters for China in operation (OECD, 2018). Note that in this study, most of the items flagged for DIF had country-specific item parameters, and the common items among the PISA participating countries were 93% on average (OECD, 2018); therefore, this strong linking ensured measurement comparability of science proficiency across different language and culture groups.

The response process data analyses, as indicated by DRT and DIA sizes in Table 7, demonstrate differences in response processes among students from different countries. The sixth column of Table 7 presents the DRT.SS effect size, an indicator of differential response time for student groups who performed at the same proficiency level. Given similar performance levels, Chinese students spent significantly longer time on Items 6, 8, 10, and 16 than Canadian students. The seventh column is the DRT.TT effect size, an indicator of differential response time for students who had similar overall testing time. Conditioning on TRT on the block, Chinese students spent significantly longer time on

Table 7. DIF, DRT, and DIA results for China (focal group) and Canada (reference group) on one block.

Item	Average percent score	Median Time (seconds)	Average No. of Actions	DIF	DRT.SS	DRT. TT	DIA.SS
1	0.41	46.29	2.17	0.35	−5.16	−10.43	0.92
2	0.8	49.86	5.45	−0.02	1.13	−3.18	2.65
3	0.67	44.45	6.11	−0.27	10.72	6.57	2.28
4	1.12	60.18	7.05	0.11	18.68	11.44	2.89
5	0.76	61.82	13.92	0.05	9.01	0.65	5.71
6	0.49	109.47	161.41	0.04	83.68	62.21	−140.30
7	1.01	71.97	15.24	−0.04	−1.31	−9.65	1.36
8	0.45	138.79	162.75	−0.04	59.33	33.66	−144.03
9	0.8	60.22	3.74	−0.02	−11.13	−18.36	2.74
10	0.41	103.91	152.89	0.04	30.31	9.47	−140.56
11	0.75	31.3	3.16	0.14	−10.29	−13.86	0.49
12	0.65	30.49	3.72	−0.11	0.46	−2.84	1.48
13	0.64	44.48	4.54	0.02	−5.94	−11.32	0.72
14	0.87	56.36	5.66	−0.24	−10.67	−17.50	1.02
15	0.33	61.47	7.12	-0.09	4.94	−4.09	2.35
16	0.42	84.52	160.22	−0.04	33.69	20.49	−150.13
17	0.73	51.96	7.88	0.09	−11.11	−16.49	1.43

Note. The bold-faced value represents the statistical significance with p-value < 0.05.

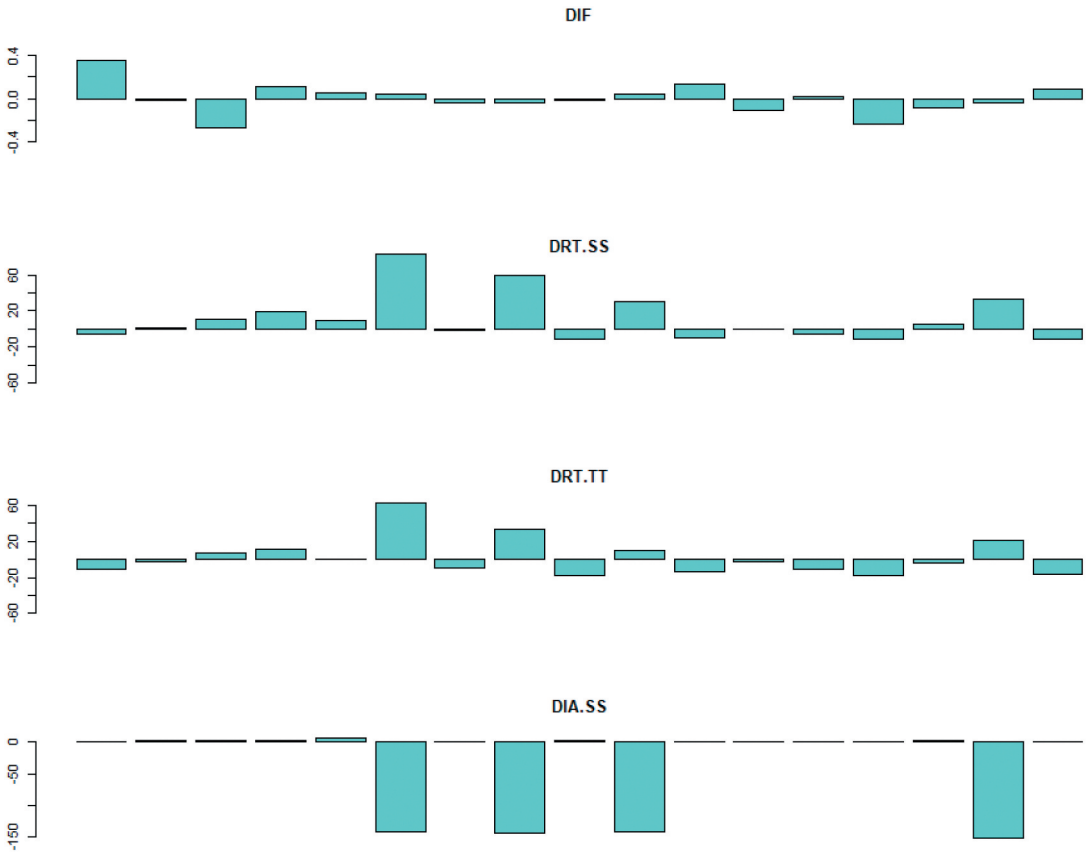

Figure 7. The effect sizes of DIF (top bar graph), DRT.SS (2rd bar graph), DRT.TT (3rd bar graph), and DIA.SS (bottom bar graph) on the first block for China and Canada comparison.

Items 6, 8, and 16 than Canadian students. The last column is the effect size of DIA.SS, an indicator of differential number of actions for students who performed at the same proficiency level. Given similar plausible values, Chinese students took much fewer actions on Items 6, 8, 10 and 16 than Canadian

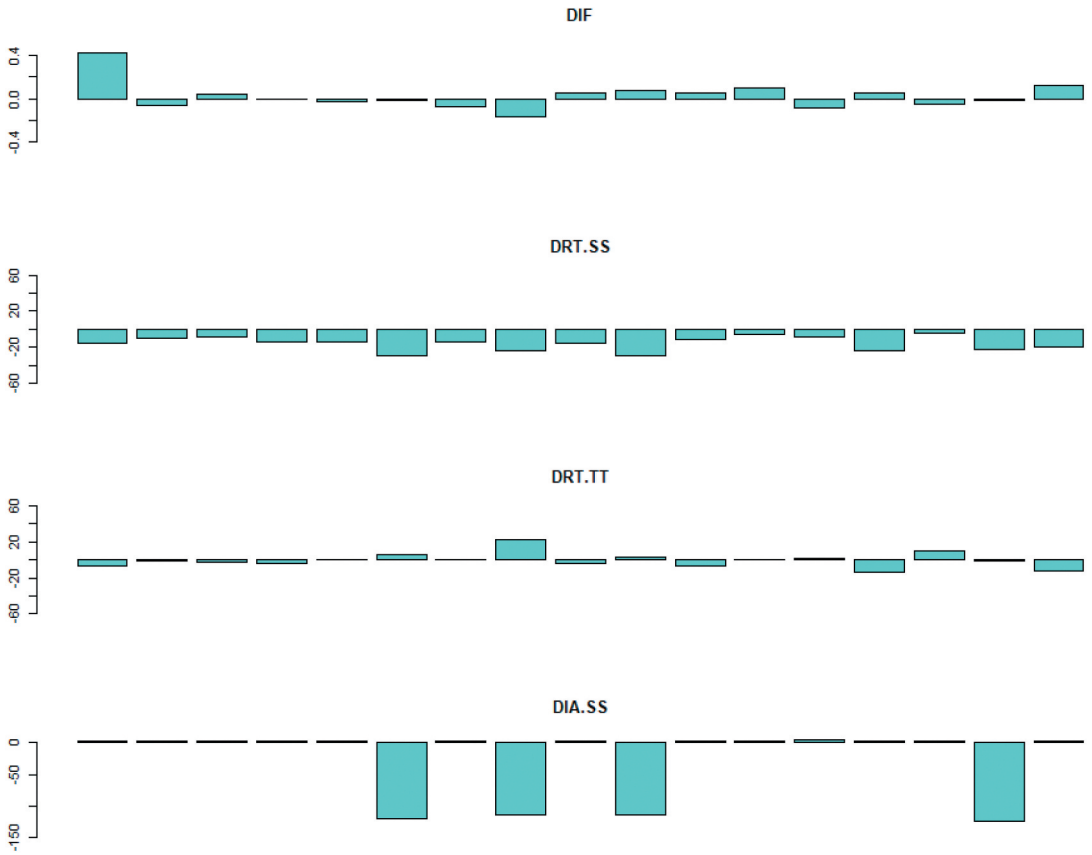

Figure 8. The effect sizes of DIF, DRT.SS, DRT.TT, and DIA.SS for Korea and Canada comparisons.

students. Figure 7 presents a visual display of the effect sizes for DIF (top bar graph), DRT.TT (second bar graph), DRT.SS (third bar graph), and DIA.SS (bottom bar graph) on this block, respectively. Figure 7 clearly shows that on the CR items (Items 6, 8, 10, 16), students' response processes were different between China and Canada – Chinese students spent much longer time but much fewer actions on these CR items.

Similar analyses were conducted for Korea vs. Canada and USA vs. Canada comparisons, respectively. Because of space limitation, we only show the effect size plots in Figures 8 and 9. Figure 8 shows the Korea vs. Canada comparison. The top bar graph shows that a few items were flagged for DIF. The second bar graph (DRT.SS) shows that given similar plausible values, Korean students spent shorter time on all items than Canadian students; but given similar TRT on the block, the third bar graph (DRT.TT) shows that Korean students spent relatively longer time on a few items than Canadian students. In addition, the bottom bar graph (DIA.SS) shows that Korean students took much fewer actions on the four CR items (Items 6, 8, 10, 16) than Canadian students. On the contrary, Figure 9 shows that DIF, DRT.TT, DIA.SS hardly flagged any items between the USA and Canada comparison, but DRT.SS shows that given similar plausible values, USA students spent slightly longer time on all items than Canadian students.

Summary of Study 2

The findings from Study 2 illustrated the use of process data in identifying timing and action differences in response process among cultural and language groups that may be overlooked by DIF

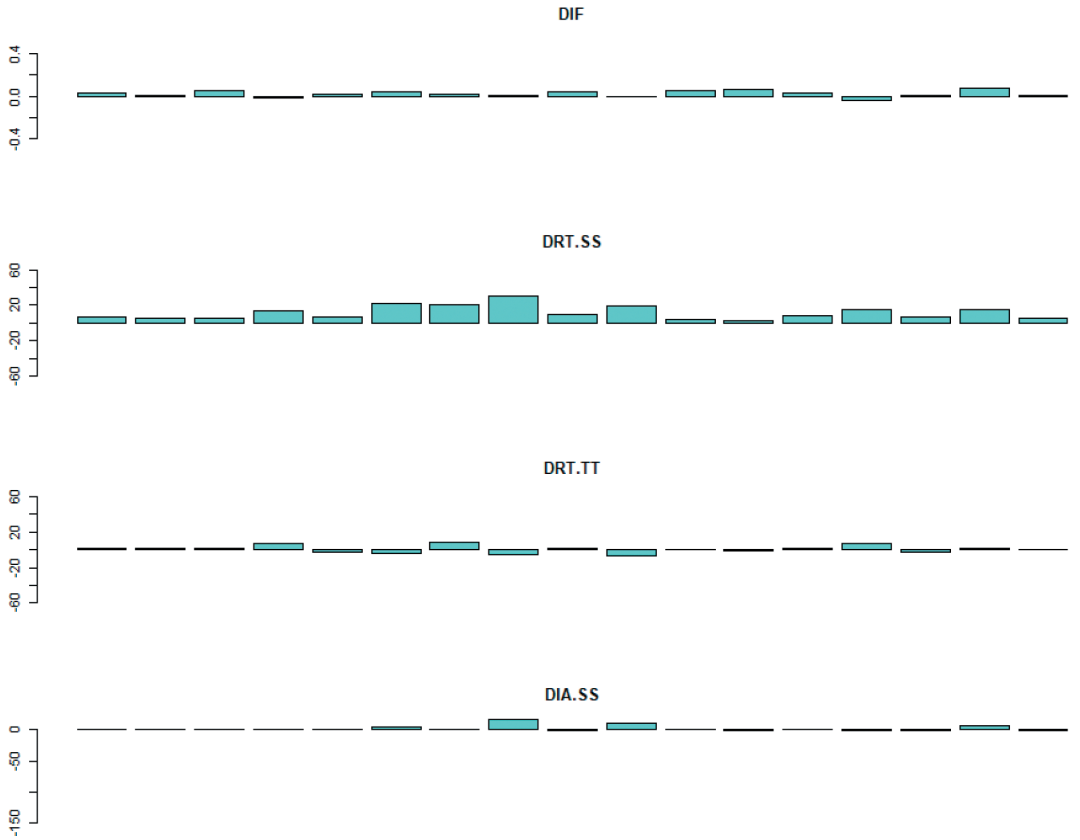

Figure 9. The effect sizes of DIF, DRT.SS, DRT.TT, and DIA.SS for USA and Canada comparisons.

analyses alone. Among the studied country groups, Canada, China, and Korea are the top performing countries in the Science assessment, and performance of China and Korea are similar (OECD, 2018, Chapter 12). However, Korea had an unusually short median response time on the assessment compared to all the other countries, including the best performing country Singapore (OECD, 2018, Chapter 9). Canada and China are comparable in terms of the median response time on the assessment, but Chinese students spent much longer time but had fewer actions on CR items than Canadian students. One hypothesis is that differences in language input methods (input method editor; IME) played a role. Chinese is a character-based language, and the commonly used Pinyin IME requires two steps to enter a Chinese word on computer: The first step uses letters on the keyboard to spell the sound of the word, and the second step is to find the corresponding characters by scanning through homophones in a pop-up list. The first step usually requires few letters (i.e. keyboard actions) to spell the pronunciation of a Chinese word than spelling an English word with the same meaning, but the second step usually takes equal amount of time, if not longer, to find the correct Chinese character (Bi, Smith, & Zhai, 2012). Because of similar input method editors for Japanese and Chinese, as expected, our extended Japan vs Canada analyses show similar results to that obtained from China vs. Canada analyses; that is, when matched on Science proficiency, Chinese and Japanese students spent longer time and had few actions on CR items than Canadian students. In contrast, Korea abandoned teaching Hanja in 1971 (Wikipedia, 2019), so Korean words are input phonetically, and words are entered in one step. Therefore, it is much more efficient to type a Korean word than a Chinese word with the same meaning, which might explain why Korean students used substantially fewer actions and shorter time on the CR items compared to

Canadian students, in addition to their possibly higher keyboarding skills. As for the USA vs. Canada comparison, we did not observe alarming differences in DIF, DRT, and DIA, probably because both groups are English language groups, even though the two countries had a large difference in Science performance.

For most of current educational assessments, test taking speed is not a skill the tests are designed to measure. However, test takers' speed and RT management unavoidably impact their performance on assessments because most of the assessments are timed (in the case of PISA Science, students need to finish the test within 60 minutes). Process data have a complex relationship with the constructs that the assessments is designed to measure. As we observed in the PISA data, relationships between item responses, response times, or actions and the construct may vary across groups. Response processes are functions of the targeted construct as well as test taking strategies, test taker's exposure to curriculum and instruction, familiarity of technology if the test is administrated on a digital environment, and cultural and language differences of the test takers. If there is evidence in process data that shows one or a few of the above-mentioned factors may have hindered students' performance on PISA assessments for certain language or cultural groups, score comparability may be compromised. For example, if students in a language group had to devote much more time on typing words because of difference in input method editors, they might lose time to work on other items on the assessment. These insights highlight the importance of examining response processes in addition to response patterns as is done in DIF analyses in investigating comparability of measurements and score meaning.

Discussion

Digital assessments provide opportunities for using new data sources for comparing groups to gain insights to inform policy and practice, as well as for examining the comparability of scores and score meaning for the comparison groups. Two types of response process data – response time and number of actions students take – are traces of response processes that can be particularly important in examining the degree to which students from different cultural and language background engage with the assessment tasks and inform inferences about comparability of measurement across groups.

In this paper we examined the possibility of using these two types of response process data, response time and number of actions, in examining measurement comparability. We proposed three statistics for examining measurement comparability: differential response time (DRT) using two different conditioning variables, total score (DRT.SS) and total response time (DRT.TT), and differential item action (DIA). We illustrated and discussed the use of these statistics in two case studies. The results in the case studies indicate that when group comparisons are restricted to students' final responses/product data using DIF analyses, significant differences in response processes for examinees from different language and cultural groups may be overlooked.

As discussed in the introduction, there are two goals of using process data: One is to help understand students' test taking strategies, improve test design, and validate final-product-based scores the test measures, and the other one is to use response process data as evidence of a targeted construct. Our two case studies show the use of process data for the first purpose. Process data analyses derived from the traditional assessments that are based on final responses and products may provide insights on why students from certain language and culture groups exhibited different test taking behaviors and whether certain groups may be disadvantaged by factors that the tests are not designed to measure. In the first study, we analyzed the ELL and non-ELL group differences on a mathematics assessment presented in the same language, and in the second one, we investigated country group differences on a science assessment translated into different languages. In Study 1, analyses of timing data indicate that, in addition to their limited proficiency on mathematics, the low performance of ELL might also be partly due to longer content processing time, lower test engagement, or both, caused by the cognitive demands for English language proficiency. More investigation is necessary to address what can be done to effectively improve ELL students' mathematics performance, including extending testing time, embedding dictionaries, audio translations, and other assistance. In Study 2, timing and

action data analyses showed that different country groups show different behaviors in responding to test questions, particularly for the constructed response items where students need to type short answers on computers. Students who used character-based language such as Chinese and Japanese used much longer time but performed much fewer actions to produce answers to CR items than Canadian students, even though these groups are comparable in overall Science proficiency and overall test response time. On the other hand, Korean and Chinese students are very similar in Science proficiency, but Korean students were much faster and more efficient in taking the test. It is important to further evaluate whether language differences have impact on measurement comparability.

Because our process data were collected from operational testing programs, findings from both case studies may need carefully designed experiments to make valid inferences about patterns we have observed in these assessments. However, insights from both case studies demonstrate the importance of using response process data in addition to response data in examining measurement comparability for groups. In addition, they provide empirical evidence on how and the degree to which ELL's test taking behaviors may be affected by their limitations in English language proficiency, and how test language might affect response time and strategies in responding test items in international assessment contexts.

There are some limitations in our studies. In the two studies presented, we used response times and number of actions, which are important variables but carrying limited information of the full spectrum of the response process as to how a test taker navigated through the assessment. The analysis methods, DRT or DIA, are more descriptive than statistical modeling. It is also worth noting that, for DIF analyses (Guo & Dorans, 2019 and reference therein), when the comparison groups differ in the matching criterion greatly, the DIF results may be different from those when the groups are similar (Oliveri & Ercikan, 2011). Further studies can investigate how group differences impact DRT results and can explore what advanced statistical and data mining methods can be used to extract inform from the response processes.

References

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

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association (AERA).

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing* (2nd ed.). Washington, DC: American Educational Research Association (AERA).

Bi, X., Smith, B., & Zhai, S. (2012). Multilingual touchscreen keyboarding design and optimization. *Human-Computer Interaction*, *27*, 352–382.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. Assessment in education. *Principles, Policy & Practice*, *5*, 7–74.

Clifton, C., Jr, Ferreira, F., Henderson, J. M., Inhoff, A. W., Liversedge, S., Reichle, E. D., & Schotter, E. R. (2016). Eye movements in reading and information processing: Keith Rayner's 40 year legacy. *Journal of Memory and Language*, *86*, 1–19. doi:10.1016/j.jml.2015.07.004

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. doi:10.1111/j.1745-3984.1986.tb00255.x

Dorans, N. J. (2013). ETS contributions to the quantitative assessment of item, test, and score fairness (Research Report No. RR-13-27, SPC-13-04). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2013.tb02334.x

Drasgow, F. (1986). Encyclopedia of statistical sciences. In N. Johnson & S. Kotz (Eds.), *Polychoric and polyserial correlations* (pp. 68–74). New York, NY: Wiley.

Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research*, *29*(6), 543–553. doi:10.1016/S0883-0355(98)00047-0

Ercikan, K. (2006). Developments in assessment of student learning and achievement. In P. A. Alexander & P. H. Winne (Eds.), *American Psychological Association, division 15, Handbook of educational psychology* (2nd ed.). Lawrence Erlbaum Associates.

Ercikan, K., & Lyons-Thomas, J. (2013). Adapting tests for use in other languages and cultures. In K. Geisinger (Ed.), *APA handbook of testing and assessment in psychology* (Vol. 3, pp. 545–569). Washington, DC: American Psychological Association.

Ercikan, K., Arim, R. G., Law, D. M., Lacroix, S., Gagnon, F., & Domene, J. F. (2010). Application of think-aloud protocols in examining sources of differential item functioning. *Educational Measurement: Issues and Practice*, *29*, 24–35. doi:10.1111/j.1745-3992.2010.00173.x

Ercikan, K., & Elliott, S. N. (2015). Assessment as a tool for communication and improving educational equity. *A white paper for the Smarter Balanced Assessment Consortium*.

Ercikan, K., & Pellegrino, J., (Eds.). (2017). *Validation of score meaning in the next generation of assessments: The use of response processes*. (An edited volume for NCME book series). Routledge.

Ercikan, K., Roth, W.-M., & Asil, M. (2015). Cautions about uses of international assessments. *Teachers College Record*, *117*, 1–28.

Ercikan, K., Yue, M., Lyons-Thomas, J., Sandilands, D., Roth, W.-M., & Simon, M. (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*, 1–23. doi:10.1080/15305058.2014.957382

Ferreira, F., & Henderson, J. M. (2004). Introduction to the interface of vision, language, and action. In J. M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (pp. ixxiv). New York, NY: Psychology Press.

Goldhammer, F., Naumann, J., & Greiff, S. (2015). More is not always better: The relation between item response and item response time in Raven's matrices. *Journal of Intelligence*, *3*, 21–40. doi:10.3390/jintelligence3010021

Guo, H., & Dorans, N. (2019). Using weighted sum scores to close the gap between DIF practice and theory. *Journal of Educational Measurement*. (in press). doi:10.1111/jedm.12258

Guo, H., Rios, J., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education*, *29*(3), 173–183. doi:10.1080/08957347.2016.1171766

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice*, *18*(4), 5–9. doi:10.1111/j.1745-3992.1999.tb00276.x

He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with N-grams: Insights from a computer-based large-scale assessment. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 749–776). Hershey, PA: Information Science Reference.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel- Haenszel procedure. In H. Waiter & H. I. Braun (Eds,) *In Test Validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hilldsdale, NJ: Lawrence Erlbaum Associates.

Lai, M., Tsai, M., Yang, F., Hsu, C., Liu, T., Lee, S., … Tsai, C. (2013). A review of using eye tracking technology in exploring learning 2000–2012. *Educational Research Review*, *10*, 90–115. doi:10.1016/j.edurev.2013.10.001

Leighton, J. P. (2017). Collecting and analyzing verbal response process data in the service of interpretive and validity arguments. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments* (pp. 39–51). New York, NY: Routledge.

Lindquist, E. F. (ed.). (1951). *Educational measurement.* Washington, DC: American Council on Education.

Liversedge, S., Schroeder, S., Hyönä, J., & Rayner, K. (2015). Emerging issues in developmental eye-tracking research: Insights from the workshop in Hannover, October 2013. *Journal of Cognitive Psychology*, *27*(5), 677–683. doi:10.1080/20445911.2015.1053487

Maddox, B., Bayliss, A. P., Fleming, P., Engelhardt, S., Gareth, E., & Borgonovi, F. (2018). Observing response processes with eye tracking in international large-scale assessments: Evidence from the OECD PIAAC assessment. *European Journal of Psychology of Education*, *33*(3), 543–558. doi:10.1007/s10212-018-0380-2

Menken, K. (2008). *English learners left behind: Standardized testing as language policy. Bilingual education and bilingualism* (pp. 65). Tonawanda, NY: Multilingual Matters Ltd.

Mislevy, R. J. & Sheehan, K. M. (1987). Marginal estimation procedures. In A. E. Beaton (Ed.), *Implementing the New Design:The NAEP* 1983–84 *Technical Report* (Report No. 15-TR-20). Princeton, NJ: Educational Testing Service.

OECD. (2018). *PISA 2015 technical report*. Retrieved from http://www.oecd.org/pisa/data/2015-technical-report/

Oliveri, M., & Ercikan, K. (2011). Do different approaches to examining construct comparability lead to similar conclusions? *Applied Measurement in Education*, *24*, 349–366. doi:10.1080/08957347.2011.607063

Oranje, A., Gorin, J., Jia, Y., & Kerr, D. (2017). Collecting, analyzing, and interpreting response time, eye-tracking, and log data. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments* (pp. 39–51). New York, NY: Routledge.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responses on aggregated scores: To filter unmotivated examinees or not? *International Journal of Testing*, *17*, 74–104. doi:10.1080/15305058.2016.1231193

Ryans, D. J., & Frederiksen, N. (1951). Performance tests of educational achievement. In E. F. Lindquist (Ed.), *Educational measurement* (1st ed., pp. 455–494). Washington, DC: American Council on Education.

Van der Linden, W. J. (2011). Test design and speededness. *Journal of Educational Measurement*, *48*(1), 44–60. doi:10.1111/j.1745-3984.2010.00130.x

Wikipedia. (2019). *Hanjo*. Retrieved from https://en.wikipedia.org/wiki/Hanja

Willner, L., Rivera, C., & Acosta, B. (2007). *Decision-making practices of urban districts for including and accommodating English language learners in NAEP –School-based perspectives*. Arlington, VA: The George Washington University Center for Equity and Excellence in Education.

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implication. *Educational Measurement: Issues and Practice*, *36*, 52–61. doi:10.1111/emip.12165

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183. doi:10.1207/s15324818ame1802_2

Zieky, M. (1993). Practical questions in the use of DIP statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Erlbaum.

Zieky, M. (2011). The origins of procedures for using differential item functioning statistics at Educational Testing Service. In N. J.Dorans & S. Sinharay (Eds.), *Looking back: Proceedings of a conference in honor of Paul W. Holland* (pp. 115–127). New York, NY: Springer.

Zwick, R. (1990). When Do Item Response Function and Mantel-Haenszel Definition of Differential Item Functioning Coincide? *Journal of Educational Statistics*, 15(3), 185–197.

Zwick, R. (2010). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (Research Report No. RR-12-08). Princeton, NJ: Educational Testing Service.

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). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2012.tb02290.x

Zwick, R., & Thayer, D. (1996). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics*, *21*(3), 187–201. doi:10.3102/10769986021003187