Proceedings of the 90th Annual International Meeting of the Psychometric Society, Minneapolis, United States (2025).

Culturally Responsive Assessments for African-American Examinees

Sandip Sinharay,* Matthew S. Johnson, and Srinivasa Pillarisetti

Educational Testing Service
*Corresponding author. Email: ssinharay@ets.org

Abstract

There are rising concerns about inequities fostered by traditional assessments (e.g., Bennett, 2023; Buzick et al., 2023; Hughes, 2023). One alternative to traditional assessments is culturally responsive assessments, which are assessments that evaluate students' knowledge, skills, and understanding in a way that takes into account their unique cultural identities (e.g., Hood, 1998; Landl, 2021). We adapt several mathematics items on a large-scale survey assessment for eighth-grade students to make them more responsive to African-American cultural contexts. Through a study conducted with several hundred students across multiple U.S. states, we evaluated the impact of cultural context on student performance and student behavior using a hierarchical approach for the joint modeling of the item scores and response times. Results revealed minimal differences in examinee performance and examinee behavior between the original and adapted test items.

Keywords: fairness; item response theory

1. Introduction

Critics have claimed that traditional or standardized assessments perpetuate social injustice through their content, design, and use (e.g., Bennett, 2023; Buzick et al., 2023; Hood, 1998; Hughes, 2023; Randall, 2021; Taylor, 2022). Such criticisms have called for the development of alternative assessments that are more fair and equitable. One set of alternatives to traditional assessments are culturally responsive assessments (CRAs), which are assessments that evaluate students' knowledge, skills, and understanding in a way that takes into account their unique cultural identities (e.g., Hood, 1998; Landl, 2021; Lee, 1998; Walker et al., 2023). Even though Bennett (2023), Walker et al. (2023), Evans and Taylor (2025) etc. developed the theory underpinning the CRAs, there is a lack of actual CRAs and of data originating from CRAs. This paper is one of the first attempts to fill that void. We developed a CRA that involves the administration of two versions of a test to two cultural groups—whites and African-Americans—and studied the difference in performance, response time, engagement, and motivation of the two groups over the two versions. The goal of the study was to examine whether cultural context affects examinee behavior or performance.

The next section—providing some background—starts with a discussion of the criticisms of traditional assessments and then includes a brief description of CRAs and of an psychometric/statistical approach suggested by Sinharay and Johnson (2024) for computing comparable scores from these assessments. In the *Development and Administrations* section, we discuss the development of a CRA for African-American students and its administration to several hundred middle-school students. The Model section first describes an existing model for the item scores and then describes a hierarchical model that we developed to jointly analyze the item scores and response times originating from the

test administration. The *Findings* section includes the results from the analysis of the data. We then provide some conclusions and discuss limitations and future works.

2. Background

2.1 Criticisms of traditional Assessments

The use of tools such as differential item functioning (DIF; e.g., Holland & Thayer, 1986), generalizability theory (e.g., Tavakol et al., 2024), measurement invariance (e.g., Millsap, 2011), and differential validity (e.g., Richardson, 1936) ensure the fairness of scores on traditional assessments. Experts have asserted that in spite of the application of these tools, traditional assessments are not completely fair and perpetuate social injustice through their content, design, and use (e.g., Sireci & Randall, 2021, p. 130). Messick (1995) and Koelsch et al. (1995) suggested that low test scores for students of color may reflect the failure of test developers to consider these students' cultural experiences in the assessments. They also stated that if the students' experiences in a cultural group are not taken into account when assessment tasks are developed and scored, the evaluation process of how well a student has learned will be flawed and that a failing performance may indicate the degree of disconnection between the task and the student's frame of reference, rather than the degree of mastery of the knowledge and skills being assessed.

These criticisms of traditional assessments imply the presence of *cultural bias* (e.g., Veale & Foreman, 1983) in traditional assessments, where cultural bias is defined as the extent to which the scores on an assessment are unfair to one or more cultural groups; the presence of cultural bias against a cultural group leads to an underestimation of the true performance of the groups' members on average.

2.2 Culturally Responsive Assessments

Several approaches have been suggested to reduce or eliminate the *cultural bias* in traditional assessments. One of them is the use of CRAs, which are assessments that evaluate students' knowledge, skills, and understanding in a way that takes into account their unique cultural identities (Bennett et al., 2025; Evans & Taylor, 2025; Landl, 2021; Ray & Sinharay, 2026; Walker et al., 2023). These assessments are supposed to promote deeper learning by design, allowing examinees to see themselves represented in material that affirms their cultural identities and perspectives. In addition, CRAs are believed to increase motivation and engagement for all examinees (Hood, 1998) and to lead to more fair and valid scores and conclusions. Sireci (2020) stated that CRAs may ensure sufficient flexibility in testing conditions to yield the most accurate measurement of proficiency for each examinee.

There are several types of assessments that are conceptually similar to CRAs (Bennett, 2025a). They include anti-racist assessment, equitable assessment, socioculturally responsive assessment, culturally sensitive assessment, culturally relevant assessment, culturally sustaining assessment, and culturally and socially responsive assessment (e.g., Bennett, 2023; Gordon, 1995; O'Dwyer et al., 2023; Randall, 2021, 2023; Taylor, 2022). These assessments have the common element that they adjust for individual differences in social, cultural, and other relevant characteristics in an attempt to lead to more fair scores, especially for traditionally underserved examinee groups. However, as described by Bennett (2025b), they differ on target population, setting, principles, goals, audiences etc., the ideas underlying them often overlap, and there is a lack of specific details about several of them.

There are a handful examples of CRAs in practice. The first is the Kaiapuni Assessment of Education Outcomes (KĀ'EO; e.g., Kūkea Shultz & Englert, 2021) that has been offered since 2016 to test-takers enrolled in Hawaiian language immersion programs in Hawaii, USA. The KĀ'EO program is a first step toward better assessing Hawaiian ways of knowing and has become a mechanism for self-determination and created a way for culture, language, and community to truly drive instruction and, as a result, assessment (Kūkea Shultz & Englert, 2023). The second is a CRA

that Dobrescu et al. (2021) developed from the National Assessment Program Literacy and Numeracy (NAPLAN) and administered in 2021 to indigenous and non-indigenous school students in 6th and 8th grade in New South Wales, Australia. They assessed two subject areas—reading and mathematics. In each grade and each subject area, the set of items administered to the indigenous students and that to the non-indigenous students had a one-to-one correspondence. Thus, for each item administered to the non-indigenous students, the set of items administered to the indigenous students included an item that is supposed to measure the same knowledge, skills, and abilities, but reflects the cultural experiences of the indigenous students. Dobrescu et al. (2021) reported that the CRA had large effects for their data, representing what they reported to be 33% of the rural-urban gap and 50% of the indigenous-non-indigenous gap, on performance in the 6th grade reading assessment.[1] Dobrescu et al. (2021) found that the CRA had a small positive effect on the performance of the indigenous examinees on the eighth-grade mathematics assessment and no effect on the sixth-grade mathematics assessment. Sinharay et al. (2025) performed a research study in which they adapted several Grade 8 mathematics items from a large-scale survey assessment (LSSA) to include vocabulary, scenarios, and contexts expected to be more familiar to Hispanic students living in the United States. The adapted test items led to a 25% reduction in the average score difference between White and Hispanic students, but little if any difference in examinee behavior indicative of motivation and engagement.

Bennett (2023, p. 90) noted that there are at least the following three types of CRAs: (a) those in which various versions a test is administered to different cultural groups; (b) those in which a test is adapted, possibly by a machine-driven algorithm, to each individual's examinees' cognitive abilities and cultural characteristics;[2] and (c) those in which examinees are given choice of problems or tools. In this paper, we consider the first of the three types.

Several experts including Bennett (2023), Hood (1998), and Walker et al. (2023) have elaborated on the principles behind and potential benefits of CRAs and other similar assessment conceptions. Hughes (2023) recommended more applications of CRAs and also recommended acknowledging the influence of sociocultural factors on test performance and conducting additional research to estimate the sizes of the effects of different sociocultural factors on the performance of racially and culturally diverse groups of test takers. Unfortunately, there is a lack of large-scale administrations of CRAs or of other similar assessment conceptions to examinees, with only the two aforementioned exceptions. Thus, it is not surprising that Bennett (2023) and Walker et al. (2023) recommended conducting empirical studies on CRAs.

3. A Culturally Responsive Assessment: Construction and Data Collection

We first chose 20 items from the Grade 8 Mathematics test belonging to a LSSA in order to develop a CRA for African-American students. The 20 items were chosen from a website that lists all publicly available items from this specific LSSA. We adapted the 20 items to align more closely with the sociocultural context of African-American students living in the Northeast USA. The adaptation process was governed by a framework for CRA that integrates research and development methods to classify levels of sociocultural responsiveness and guides the item adaptation process to align with the various levels of sociocultural responsiveness.

Figure 1 provides a graphical description of the framework, which was inspired by guidelines from Bennett (2023) and Walker et al. (2023). In the figure, "surface-level CRA" refers to CRAs whose items are obtained by making only superficial changes to items on a traditional test and "deep-level CRA" refers to CRAs whose items are obtained by making deeper (than superficial) changes to items on a traditional test.

We involved inputs from middle-school teachers, mathematics content experts who work with

[1] They reported the estimated "average treatment effect" (ATE), which is the average score on the CRA minus that on the traditional test after adjusting for the differences in covariates like socioeconomic status and previous standardized test scores. The ATE was 1.5 on the raw score scale (the maximum raw score is 33), which is about 50% of the difference in average

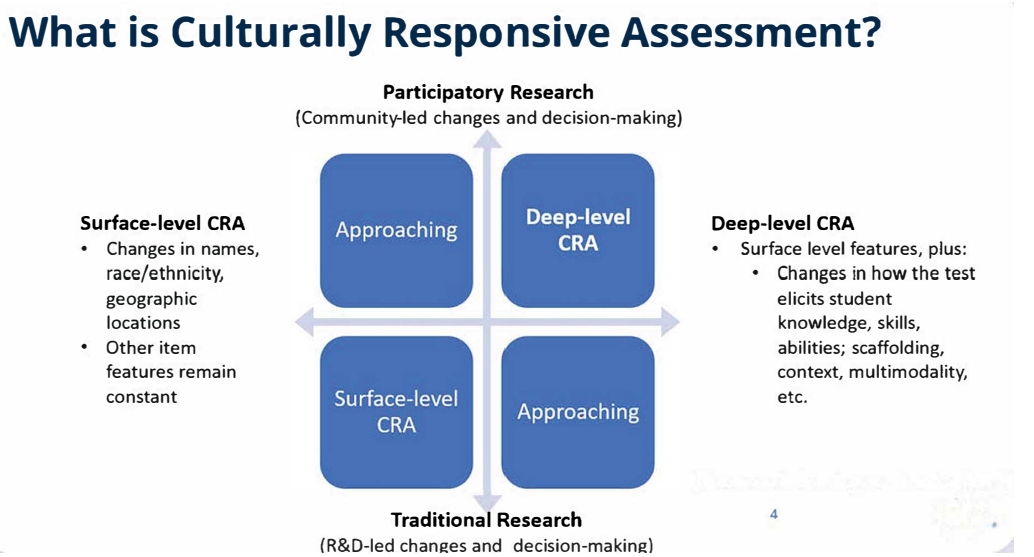

Figure 1. The framework used for item adaptation.

the LSSA that we considered, and test developers with African-American backgrounds during the item adaptation process. Fairness review was performed of the adapted items by two internal groups at our institution.

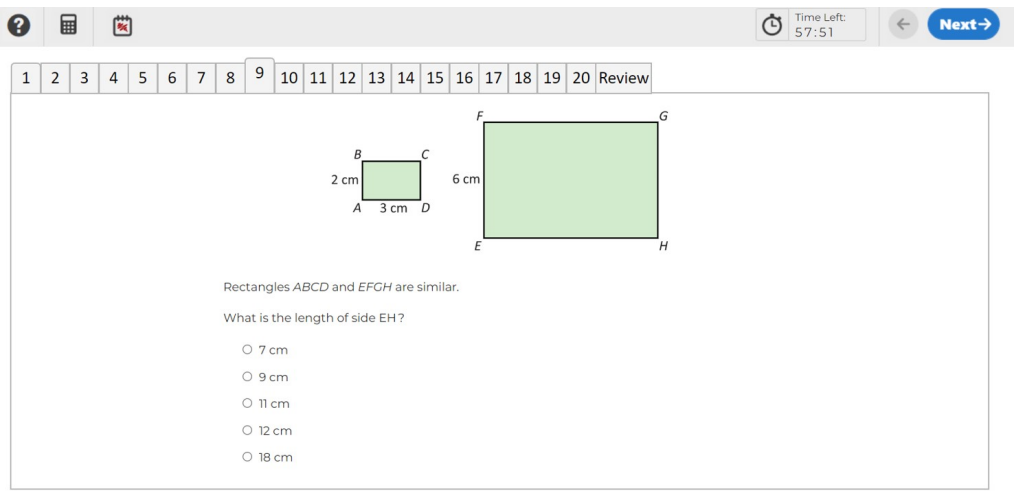

Figure 2. Original Item 9.

Figure 2 shows an original item from the LSSA (Item 9 of the test) while Figure 3 shows the adapted version of the item. The appendix shows another original item (Item 14) and its adapted version.

There are several available approaches for developing CRAs. One approach involves implement-

scores of the non-indigenous students and indigenous students.

[2]This type of CRAs are also referred to as personalized assessments (e.g., Bennett, 2024; Buzick et al., 2023).

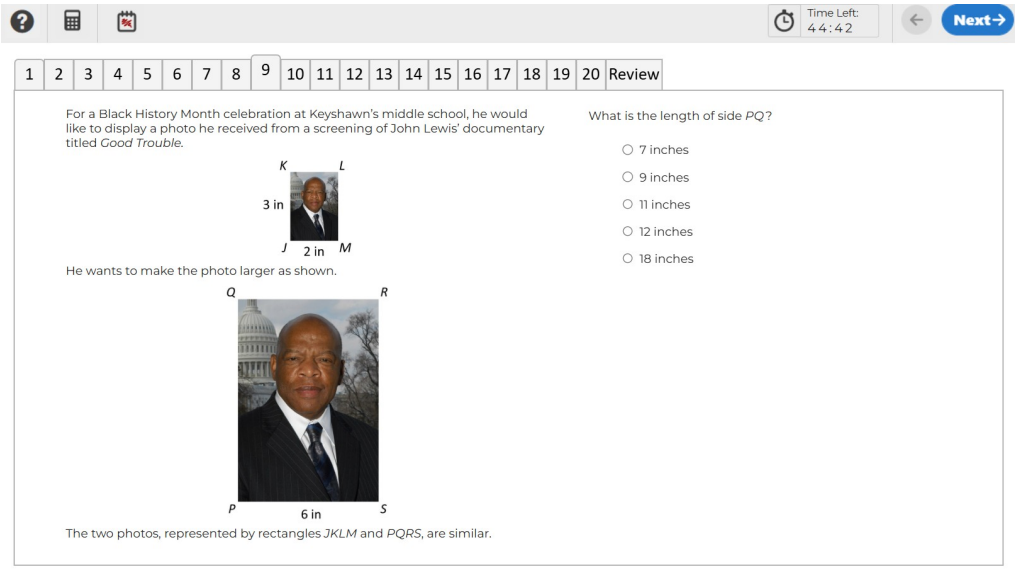

Figure 3. Adapted Version of Item 9.

ing the Principle 1 of Bennett (2023) by crafting item contexts that depict members of a specific demographic group or, more preferably, appeal to the experiences of specific demographic groups. In addition, Evans and Taylor (2025, pp. 5-6) noted that CRAs may use contexts or scenarios that students may find familiar, interesting, or engaging (e.g., music, art, sports) and tasks where students see themselves, their communities, and cultures represented in some form or fashion, with this representation and exposure sometimes focusing on surface-level features of culture (e.g., food, fashion, festivities). Thus, we believe that the assessment that we developed qualifies as a CRA for African-American students.

We constructed two test forms. The 20 original items from the LSSA (referred to as "original items 1-20") and their adapted counterparts ("adapted items 1-20") comprised Forms 1 and 2, respectively. All items were dichotomously scored. The two forms of the CRA were administered to eighth-grade students in a school located in Louisiana in Spring and Fall of 2025.[3] The total sample size was approximately 500. The number of white, African-American, and Hispanic examinees[4] in the sample were 150, 187, and 165, respectively. The school volunteered to participate in the study and received $25 for each student that participated in our study. We randomly assigned the two forms to the examinees. Each examinee had to answer the 20 mathematics questions and complete a brief post-test questionnaire within a class period, which typically lasts 45 minutes.

4. A Hierarchical Modeling Approach for Joint Modeling of Item Scores and Response Times

4.1 A Multidimensional Item Response Theory Model for the Item Scores

A procedure based on multidimensional item response theory (IRT) modeling was suggested by Sinharay and Johnson (2024) for analyzing item scores originating from CRAs. The procedure allows for the computation of comparable scores on various versions of a CRA. Let us consider the case—one that applies to our real data example—where a CRA is administered to an examinee population that includes two distinct population groups referred to as the focal and reference groups that differ vastly in their cultural experiences. For example, the focal and reference groups could be

[3]Thus, two different cohorts of students were tested in the school.

[4]The ethnicity of each student was determined based on a student roster sent by the school administration.

indigenous and non-indigenous examinees, respectively, as in Dobrescu et al. (2021). For simplicity, each group will be assumed culturally homogeneous. Let us refer to the test forms targeted to the focal group and the reference group as the focal-group version and reference-group version (RGV), respectively. Let us also assume that for each item in the reference-group version, there exists an item in the focal-group version that is supposed to measure the same knowledge, skills, and abilities as the former but reflects the cultural experiences in the focal group.

Let $X_{ij\nu}$ denote the dichotomous score of examinee i on item j of version ν. Let $g(i)$ denote the population group to which examinee i belongs. The value of ν is 1 and 2 for focal-group version and reference-group version, respectively, and that of $g(i)$ is 1 and 2 depending on whether $g(i)$ is the focal group or the reference group, respectively. To analzye such data, Sinharay and Johnson (2024) suggested using a version of the two-dimensional between-items Rasch model (e.g., Adams et al., 1997; Reckase, 2009, p. 93) given by

$$\text{logit}(P(X_{ij\nu} = 1)) = \begin{cases} \theta_i - b_{j\nu g(i)} \text{ if } \nu = g(i), \\ \eta_i - b_{j\nu g(i)} \text{ if } \nu \neq g(i), \end{cases} \tag{1}$$

where $\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$, $b_{j\nu g(i)}$ is the true difficulty parameter of item j of version ν for group $g(i)$, θ_i is the true ability of examinee i to solve problems that are situated in the context of the examinee's cultural group, and η_i is the true ability of examinee i to solve problems that are not in the context of the examinee's cultural group. The θ_i and η_i are defined as the *within-context ability parameter* and *out-of-context ability parameter*, respectively. Thus, the probability of a correct answer is governed by θ_i if the CRA version matches the examinee's group, but by η_i if the CRA version does not match the examinee's group. The parameters b_{j11} and b_{j22} denote item difficulties within context and are referred to as *within-context difficulty* parameters. The parameters b_{j12} and b_{j21} are referred to as *out-of-context difficulty* parameters. The ability parameter vector for examinee i, $(\theta_i, \eta_i)'$, is assumed to follow a bivariate normal population distribution whose parameters vary over the examinee groups. To ensure identifiability of the parameters, the means of the bivariate normal distribution are assumed to be equal to zero for the focal group.

One has to make certain assumption(s) in order to be able to perform some form of linking (e.g., Kolen & Brennan, 2014, p. 191) that would produce scores that are comparable over the two versions of the CRA. One reasonable assumption[5] is that some within-context item parameters are equal over the focal-group version and reference-group version, or $b_{j11} = b_{j22}$, for a subset of items on the CRA. The assumption is equivalent to the assumption that for some item pairs, the item belonging to the focal-group version measures the same skill for the focal group in the same way as measured for the reference group by its counterpart belonging to the reference-group version, or, some items are common between the two versions. The set of common items can be found using expert opinion and DIF analysis, as recommended for multilingual/translated tests by Budgell et al. (1995), Ercikan and Lyons-Thomas (2013), Sireci (1997), and Sireci et al. (2016).[6]

The model provided in Equation 1 can be fittted using standard software packages that can be used to fit multidimensional IRT models. For example, one can use the multiGroup function in the R package *mirt* (Chalmers, 2012) to fit the model.

4.2 A Hierarchical/Joint Multigroup Model and its Estimation

Because the response times of the examinees on the test items were available and the aforementioned model of Sinharay and Johnson (2024) does not incorporate response times, we developed a hierarchical model to analyze the item scores and response times simultaneously. The underlying idea of the

[5]The assumption is similar to one typically made to analyze data from translated tests (e.g., Ercikan & Lyons-Thomas, 2013; Sireci, 1997; Sireci et al., 2016).

[6]For examples of such DIF analysis for real data from translated tests, see, for example, Angoff and Cook (1988) and Budgell et al. (1995).

joint model in the same as that underlying the hierarchical modeling approach that van der Linden (2007) suggested for joint modeling of item scores and response times. In the first stage of the modeling approach,

- we assumed that the item scores $X_{ij\nu}$ follow the multidimensional IRT model proposed by Sinharay and Johnson (2024).
- we assumed that the response times follow the lognormal model for response times (van der Linden, 2006), which, in our context, is given by

$$y_{ij\nu} \sim \mathcal{N}(\beta_{j\nu} - \tau_i, \frac{1}{\alpha_{j\nu}^2}), \tag{2}$$

  where $y_{ij\nu}$ is the logarithm of the response time of examinee i on item j of version ν, τ_i is the *speed parameter* of examinee i and $\alpha_{j\nu}$ and $\beta_{j\nu}$ respectively are the *time-discrimination parameter* and *time-intensity parameter* of item j of version ν.

A preliminary examination of the summary statistics of the response times revealed that any given examinee group took the same time on average on original and adapted version—that is why we did not choose two different speed parameters (one for within-context and another for out-of-context items) for the examinees. In the second stage of the modeling approach, we assumed the examinee parameters—θ_i, η_i, and τ_i—to jointly follow a trivariate normal population distribution. Because the examinees belong to two groups, a multigroup extension of the hierarchical model was needed, which was performed by assuming that the moments of the population distribution differ over the groups, that is,

$$(\theta_i, \eta_i, \tau_i)' \sim \mathcal{N}_3(\boldsymbol{\mu}_{g(i)}, \Sigma_{g(i)}), \tag{3}$$

where $\boldsymbol{\mu}_{g(i)}$ and $\Sigma_{g(i)}$ respectively are the mean vector and variance matrix for group g_i. To ensure identifiability of the group parameters, all elements of $\boldsymbol{\mu}_{g(i)}$ are assumed to be equal to 0 and all diagonals of $\Sigma_{g(i)}$ are assumed to be equal to 1 for the African-American examinee group.[7] The means and variances for the white examinee group can be estimated from the data. The model was fitted to our data set using the Hamiltonian Monte Carlo (HMC) algorithm (Gelman et al., 2014, p. 300) implemented in the rstan package (Stan Development Team, 2025) after assuming weakly informative prior distributions on the model parameters.

5. Findings

In this section, we discuss the our findings, both from the summary statistics and from the application of the aforementioned hierarchical multigroup model, from the data originating from the administration of the CRA. Results from the analyses of test scores are discussed first, followed by those from the analyses of response times, and those on motivation and engagement of the examinees.

Table 1 show the average raw scores (out of a maximum of 20), standard deviation (SD) of the raw scores, and the number of examinees that formed the basis of these calculations (n) for the African-American and white examinees on the two forms. The scores of the African-American were considerably lower than that of white examinees, which is in agreement with official results from well-known LSSAs.[8] Table 1 shows that the average scores on the adapted items (in Form 2) are lower than those on the the original items (in Form 1) for both the examinee groups. However, the difference in the averages between Forms 1 and 2 is not statistically significant (in two-sample t-tests) for either of the examinee groups.

[7] This choice does not affect the estimate of the group difference.

[8] For example, the average Grade 8 Mathematics scores on the National Assessment of Eduational Progress (NAEP) were 284 and 252, respectively, in 2022 for white and African-American examinees.

Table 1. Summary of the Raw Scores on the Two Test Forms by Subgroup.

Test	White Examinees			African-American Examinees		
Form	Average	SD	n	Average	SD	n
1	10.6	3.8	77	6.6	3.5	88
2	10.1	4.4	73	6.4	3.4	99

Form 1: Original items; Form 2: Adapted items; n=Sample size

We then fitted the hierarchical multigroup model described in Equations 1, 2, and 3 to compute comparable scores for the examinees. Test developers who worked on the adaptation testified to the construct equivalence of the original and adapted version. Analysis for DIF, where the two examinee groups were the white examinees who answered the original items and the African-American examinees who answered the adapted items, showed the presence of DIF on a very few items—so measurement equivalence also seems to hold for our data. We applied the following three recently developed DIF detection methods: (a) regularization-based DIF-detection approach (e.g., Wang et al., 2023); (b) DIF via robust scaling (Halpin, 2024); and (c) DIF under a multiple indicators and multiple causes (MIMIC) model (Chen et al., 2023)[9] and the results from the three methods were very close. Four out of 20 items with the least amount of DIF are assumed "common" between the original and adapted versions—so these items are assumed to measure the mathematical skills in the same way on the original and adapted items. These "common" items (Items 6, 12, 13, and 16), along with the hierarchical model described in Equations 1 and 2, were used to estimate the within-context ability (θ_i) for each examinee. To fix the location and scale, the mean and the variance of the population ability distribution of the African-American examinees were set equal to 0 and 1, respectively.[10]

Table 2. Mean Abilities.

	African-American Examinees	White Examinees
CRA (within-context)	0.00	1.11
Traditional	0.00	1.13

The averages of the estimated within-context abilities of the African-American examinees and the white examinees are shown in the first row of numbers in Table 2. The estimated group difference from the CRA, which is the average of the within-context abilities for the white examinees minus the average of the within-context abilities for the African-American examinees, was 1.11, which is equivalent to about 3.78 on the raw score scale.[11] If one does not use the CRA and instead computes the group difference only from the original LSSA items (that is, from Form 1) in a multigroup IRT calibration,[12] the averages of the estimated abilities of the two groups of examinees are shown in

[9]Halpin (2024) noted that the usual assumptions made in DIF analysis are not viable in cross-cultural contexts, which often involves the development of new assessments for use in new populations. For example, it cannot be assumed that DIF is limited to only a small proportion of items on an assessment or that a subset of items without DIF ("anchors") can be reliably identified ahead of time. Therefore, we resorted to more recent DIF methods that have been found to be more robust to violations of assumptions made in traditional DIF analysis.

[10]This choice does not affect the conclusions on group differences.

[11]This conversion was found by assuming that a difference of 1 on the ability scale is equal to a difference of 3.4, which is the SD of the African-American examinees on Form 2 in Table 1, on the raw score scale.

[12]Where the mean and the variance of the population ability distribution of the African-American examinees are set equal to 0 and 1, respectively.

the second row of numbers in Table 2. The estimated group difference is 1.13, which is equivalent to about 3.84 on the raw score scale. Thus, the CRA reduces the estimated group difference by a negligible amount, that is, does not eliminate much of the large group difference, which is somewhat counter to what the proponents of CRA assert. However, this finding is similar to the finding of no effect of CRAs on Mathematics tests in Dobrescu et al. (2021).

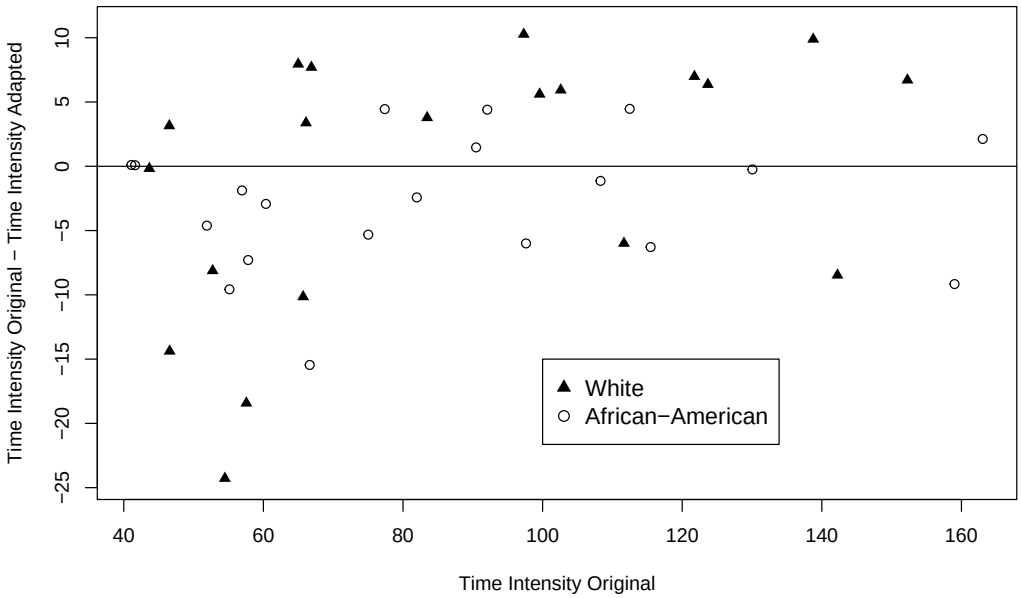

Figure 4. Plots of the differences of the estimated time intensity parameters on the original items and the adapted versions for both examinee groups.

We next proceeded to examine the results on the response times to find out if some items or their adapted versions were more or less time-consuming to either the white or African-American examinees. Figure 4 shows the differences in the estimated time intensity parameters (β_{jv}'s in Equation 2) on the original items and the adapted versions for the white examinees (solid triangle) and the African-American examinees (hollow circle) for Items 1-20. A horizontal line is drawn at the value zero; points lying around this line would indicate that the corresponding examinee group takes the same time on average to respond to the original items and adapted items. The estimates for the original items is larger than that for the adapted items for 12 items for the white examinees. However, the differences for a few items—corresponding to the bottom left part of the plot—are negative and larger in absolute value. The estimates for the adapted items are larger than that for the adapted items for 13 items for the African-American examinees. Thus, there is a tendency for the adapted items to be more time-consuming compared to the original items for the African-American examinees. However, formal hypothesis tests,[13] whose results are not reported, did not find this effect to be statistically significant.

Figure 4 compares, between the original and adapted versions, what can be thought of one-

[13]We performed (a) item-by-item z-tests comparing the two time-intensity estimates, like in an IRT differential item functioning analysis, adjusted for multiple comparison and also (b) a two-sample t-test that the average of the differences is zero.

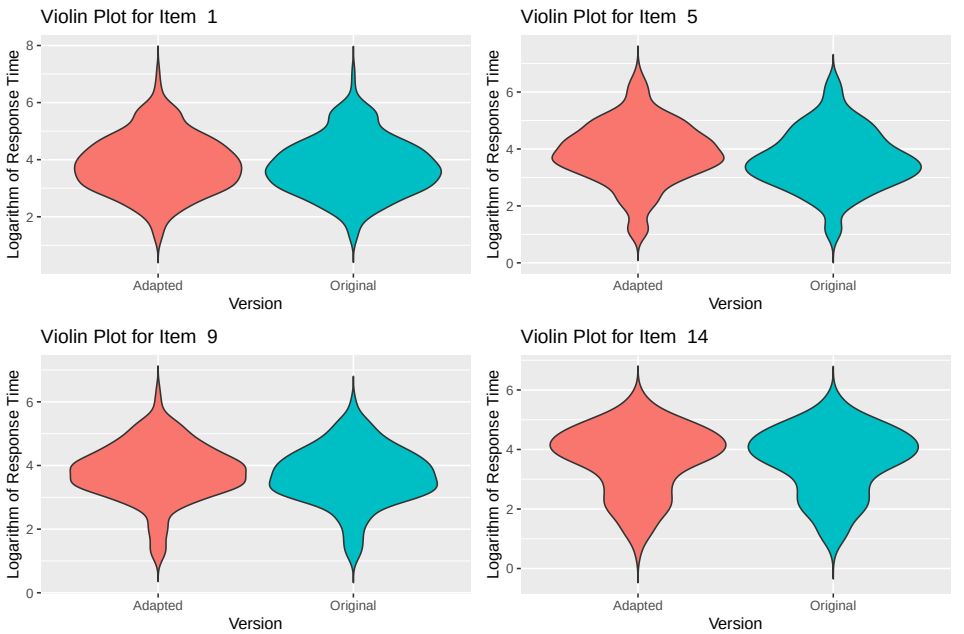

Figure 5. Violin plots showing the distributions of the logarithm of response times (in seconds) on four original items and their adapted versions for the African-American examinees.

number summaries of the response times for the items. However, it is possible to dig deeper and compare the distributions of the response times or their logarithms on the original and adapted versions for each examinee group. Figure 5 shows violin plots (e.g., Hintze & Nelson, 1998; Wright, 2019) for the distributions of the logarithms of the response time for the African-American examinees on four original items (Items 1, 5, 9, and 14) and their adapted counterparts. These items were chosen because their distributions are representative of the distributions for the full set of 20 items. A violin plot combines a box plot and a kernel-density plot into one plot. These plots show the distribution of the data with a rotated smoothed curve and serve a similar role to boxplots for showing the main features of distributions and differences among distributions. Figure 5 shows that the distributions over the two versions are similar for each item.

Bennett (2023) and Walker et al. (2023) expected traditionally underserved examinee groups to be more motivated and engaged while taking CRAs. We performed two types of analyses to examine if the African-American examinees were indeed more motivated and engaged on the adapted versions of the original LSSA items.

The first type of analyses was performed using the response times. We applied the approximation-based normative threshold method (Wise, 2017) to find the responses that can be considered to show rapid-guessing behavior. According to the method, a response time that is less than 10% of the mean response time on an item represents a rapid-guessing behavior.[14] Figure 6 shows the proportion of examinees showing rapid-guessing behavior on the original items (solid triangles) and the adapted items (hollow triangles) for the white examinees and those for the African-American examinees (solid and hollow circles, respectively). The X-axis corresponds to item number and the Y-axis corresponds to the proportion. In each panel, the proportion showing rapid-guessing behavior increases as the test progresses, even though the proportion does not reach too high. For either group, the difference in proportion showing rapid-guessing behavior between the original and adapted versions is not

[14] These 10% thresholds ranged between 4.1 and 14.4 seconds for our data.

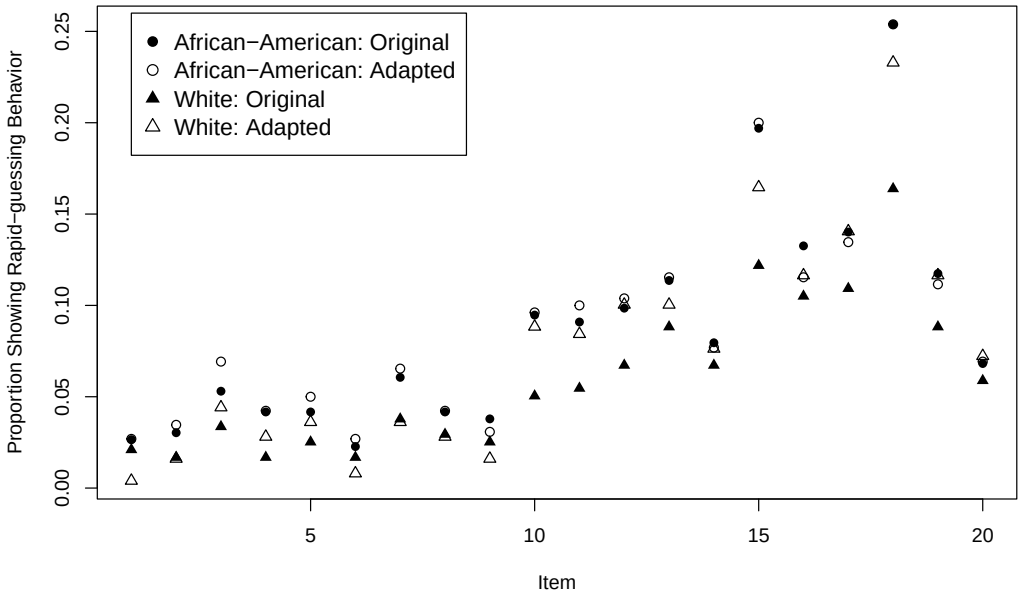

Figure 6. The proportion of examinees showing rapid-guessing behavior on the original items and their adapted versions for the white examinees and African-American examinees.

statistically significant for any item. Thus, the response times do not point to any difference in motivation between the original and adapted items for either examinee group.

The second type of analyses (regarding motivation and engagement) was performed with responses to a 8-item questionnaire that each examinee was administered after providing answers to the mathematics items. The questionnaire included several questions that were intended to inform about the level of engagement and motivation of the examinees. One question was "How enjoyable did you find completing the Math questions?", with response options (a) very enjoyable, (b) somewhat enjoyable, (c) Neutral; neither enjoyable nor unenjoyable, (d) not enjoyable, (e) very unenjoyable. Another question was "How relevant were the examples in the questions?", with response options (a) very relevant, (b) somewhat relevant, (c) Neutral; neither relevant or irrelevant, (d) somewhat irrelevant, (e) very irrelevant. Figure 7 shows the proportions of African-American examinees who chose the various options regarding whether the test questions were enjoyable (left panel) and relevant (right panel) using stacked bar charts (e.g., Nussbaumer Knaflic, 2015, p. 156). Each stacked bar chart shows the percentage choosing each answer option among African-American examinees who received the original LSSA items (left bar in each panel) or their adapted versions (right bar). There are not much differences between the proportions listed under "Original" and those listed under "Adapted" for either question and a χ^2 test of equality of the two sets of percentages[15] in Figure 7 produced non-significant results for both questions (χ^2_4 = 4.07 and 9.04, respectively; p = 0.40 and 0.06, respectively). Thus, the responses to the questionnaire do not point to any difference in motivation and engagement between the original and adapted items for either examinee group.

[15]The test was performed using the function "chisq.test" in the R software (R Core Team, 2024).

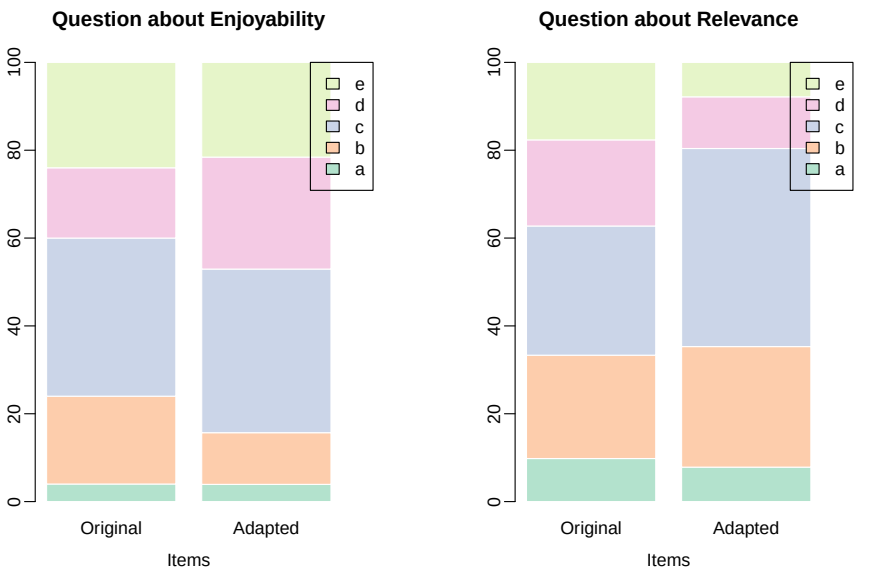

Figure 7. The Proportions (among African-American examinees) who chose the different options for the survey questions on whether the examinees found the test questions enjoyable and relevant.

6. Conclusions, Limitations, and Future Directions

This paper represents one of the first attempts to develop and administer a CRA, a type of assessment that is gaining attention because of its promise in increasing validity and reducing cultural bias for diverse populations, for African-American examinees. We choose 20 Mathematics Grade 8 items from a LSSA and then adapted them to the sociocultural context of African-American examinees residing in Northeast USA. We created two forms from the 40 items (20 original items and their adapted versions) and administered them to about 500 middle-school students in 4 schools in the Northeast USA. The analyses of the data originating from the test administration do not show much performance difference between the original items and their adapted versions for either white or African-American examinees. Application of a hierarchical multigroup model for item scores and response times reveals that the group difference estimated from the CRA is about the same as the group difference estimated only from the LSSA items.

The results from summary statistics and the hierarchical model do not show much difference with respect to response times, motivation, and engagement over the original items and their adapted versions for either examinee group. Thus, cultural context does not seem to have much impact on student performance or engagement, based on these data. This study underscores the complexity of CRAs and the necessity for continued research to maximize their potential benefit for diverse student populations.

Helms-Lorenz et al. (2003) and Malda et al. (2010) noted that in studies of cross-cultural group differences, one of two hypotheses typically holds. The first is the *Spearman hypothesis* (e.g., Jensen, 1985) that predicts one examinee group to score consistently higher than the other on both test versions, apart from random fluctuations. The second is the *cultural complexity hypothesis* (e.g., Malda et al., 2010) that predicts one examinee group to score at least as high as the other group(s) on the test version that is aligned with its own culture. Our data provides more support for the Spearman's hypothesis than the cultural complexity hypothesis.

There are several potential reasons of a lack of much performance difference between the original and adapted items in our study, often pointing to potential topics for future research. First, it is possible that in spite of the claims of reduced performance difference on CRAs, the type of CRA that we considered may not reduce group difference on a subject like mathematics where the cognitive steps students must take to solve the problems are not influenced much by the context. Note that Dobrescu et al. (2021) did not find a significant performance difference for Australian school students over indigenous and non-indigenous versions of CRAs in the subject area of mathematics. Second, it is possible that we were not able to adapt the items adequately, so that the adaptations were too superficial, to show a significant performance difference (or motivation or engagement). Third, the culture within African-American examinees differs depending on the countries from which they or their family came from (Caribbean countries or African countries etc.) or depending on whether the examinees live in a rural, suburban, or urban areas. Consequently, it is possible that a large proportion of African-American participants of our study did not find the cultural context (in our adapted items) representing their culture. One way to find out more about the reason would be to perform a larger study with a more diverse sample of African-American examinees, possibly belonging to more schools and more states, from various cultural and geographic backgrounds and stratify the examinees based on these backgrounds before administering the test forms. Fourth, it is possible that some adapted items (like item 9 shown earlier) involved an additional language load due to the adaptation process, which may have adversely affected the performance of the African-American students. Finally, as Sinharay and Johnson (2024) discussed, the difficulty difference (between the original and adapted items) and group difference (between white and African-American examinees) are confounded and no statistical analysis can disentangle the two effects for our data. It is possible to try to elicit information about the two effects from test developers and domain experts belonging to both the minoritized and reference groups in such cases.

Acknowledgement

The authors would like to thank Daniel McCaffrey and Brian Junker for their helpful comments.

Funding Statement: The research was partially supported by the National Science Foundation Award 2243041.

Competing Interests: None

References

Adams, R. J., Wilson, M. R., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1–23.

Angoff, W. H., & Cook, L. L. (1988). *Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test* (ETS Research Report Series Number 88-3). ETS. Princeton, NJ. https://doi.org/10.1002/j.2330-8516.1988.tb00259.x

Bennett, R. E. (2023). Toward a theory of socioculturally responsive assessment. *Educational Assessment*, *28*, 83–104. https://doi.org/10.1080/10627197.2023.2202312

Bennett, R. E. (2025a). A descriptive review of culturally responsive, socioculturally responsive, and related assessment conceptions. In R. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy*. Routledge/Taylor & Francis.

Bennett, R. E. (2025b). Rethinking equity and assessment through opportunity to learn. *Assessment in Education: Principles, Policy and Practice*, *32*, 5–32. https://doi.org/10.1080/0969594X.2025.2462549

Bennett, R. E., Darling-Hammond, L., & Badrinarayan, A. (Eds.). (2025). *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy*. Routledge/Taylor & Francis. https://doi.org/10.31094/2020/1

Bennett, R. E. (2024). Personalizing assessment: Dream or nightmare? *Educational Measurement: Issues and Practice*, *43*(4), 119–125. https://doi.org/10.1111/emip.12652

Budgell, G. R., Raju, N. S., & Quartetti, D. A. (1995). Analysis of differential item functioning in translated assessment instruments. *Applied Psychological Measurement*, *19*, 309–321. https://doi.org/10.1177/014662169501900401

Buzick, H. M., Casabianca, J. M., & Gholson, M. L. (2023). Personalizing large-scale assessment in practice. *Educational Measurement: Issues and Practice*, *42*(2), 5–11. https://doi.org/https://doi.org/10.1111/emip.12551

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chen, Y., Li, C., Ouyang, J., & Xu, G. (2023). DIF statistical inference without knowing anchoring items. *Psychometrika*, *88*(4), 1097–1122. https://doi.org/http://dx.doi.org/10.1007/s11336-023-09930-9

Dobrescu, I., Holden, R., Motta, A., Piccoli, A., Roberts, P., & Walker, S. (2021). *Cultural context in standardized tests* (tech. rep.). Discussion paper 2021-08, School of Economics, The University of New South Wales. Available at http://dx.doi.org/10.2139/ssrn.3983663. New South Wales, Australia.

Ercikan, K., & Lyons-Thomas, J. (2013). Adapting tests for use in other languages and cultures. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, vol. 3: Testing and assessment in school psychology and education.* (pp. 545–569). American Psychological Association. https://doi.org/10.1037/14049-026

Evans, C. M., & Taylor, C. S. (2025). *Culturally responsive assessment in classrooms and large-scale contexts*. Routledge. https://doi.org/10.4324/97810033922171

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis*. Chapman; Hall.

Gordon, E. W. (1995). Toward an equitable system of educational assessment. *The Journal of Negro Education*, *64*(3), 360–372. https://doi.org/https://doi.org/10.2307/2967215

Halpin, P. F. (2024). Differential item functioning via robust scaling. *Psychometrika*, *89*(3), 796–821. https://doi.org/http://dx.doi.org/10.1007/s11336-024-09957-6

Helms-Lorenz, M., Van de Vijver, F. J., & Poortinga, Y. H. (2003). Cross-cultural differences in cognitive performance and spearman's hypothesis. *Intelligence*, *31*(1), 9–29. https://doi.org/http://dx.doi.org/10.1016/s0160-2896(02)00111-3

Hintze, J. L., & Nelson, R. D. (1998). Violin plots: A box plot-density trace synergism. *The American Statistician*, *52*(2), 181–184. https://doi.org/http://dx.doi.org/10.2307/2685478

Holland, P. W., & Thayer, D. T. (1986). *Differential item functioning and the mantel-haenszel procedure* (ETS Research Report No. RR-86-69). ETS. Princeton, NJ.

Hood, S. (1998). Culturally responsive performance-based assessment: Conceptual and psychometric considerations. *The Journal of Negro Education*, *67*(3), 187–196. https://doi.org/10.2307/2668188

Hughes, G. B. (2023). *Improving equitable measurement and reporting in NAEP* (White Paper). AIR. https://www.air.org/resource/report/white-paper-improving-equitable-measurement-and-reporting-naep.

Jensen, A. R. (1985). The nature of the black–white difference on various psychometric tests: Spearman's hypothesis. *Behavioral and Brain Sciences*, *8*(2), 193–219. https://doi.org/http://dx.doi.org/10.1017/s0140525x00020392

Koelsch, N., Estrin, E., & Farr, B. (1995). Guide to developing equitable performance assessments [ERIC Document Reproduction Service No. ED 397-125].

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking* (3rd). Springer.

Kūkea Shultz, P., & Englert, K. (2021). Cultural validity as foundational to assessment development: An indigenous example. *Frontiers in Education*, *6:701973*. https://doi.org/10.3389/feduc.2021.701973

Kūkea Shultz, P., & Englert, K. (2023). The promise of assessments that advance social justice: An indigenous example. *Applied Measurement in Education*, *36*(3), 255–268. https://doi.org/10.1080/08957347.2023.2222031

Landl, E. (2021, October). *Culturally responsive assessment and balanced assessment systems*. REL Pacific Webinar. https://www.youtube.com/watch?v=rT3tHRyCMhA

Lee, C. D. (1998). Culturally responsive pedagogy and performance-based assessment. *The Journal of Negro Education*, *67*(3), 268–279. https://doi.org/10.2307/2668195

Malda, M., van de Vijver, F. J., & Temane, Q. M. (2010). Rugby versus soccer in South Africa: Content familiarity contributes to cross-cultural differences in cognitive test scores. *Intelligence*, *38*(6), 582–595. https://doi.org/10.1016/j.intell.2010.07.004

Messick, S. (1995). Standards of validity and the validity of standards in performance asessment. *Educational Measurement: Issues and Practice*, *14*(4), 5–8. https://doi.org/10.1111/j.1745-3992.1995.tb00881.x

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.

Nussbaumer Knaflic, C. (2015). *Storytelling with data: A data visualization guide for business professionals*. Wiley.

O'Dwyer, E. P., Sparks, J. R., & Olah, L. N. (2023). Enacting a process for developing culturally relevant classroom assessments [doi=10.1080/08957347.2023.2214652]. *Applied Measurement in Education*.

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Randall, J. (2021). "Color-Neutral" is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice*, *40*(4), 82–90. https://doi.org/10.1111/emip.12429

Randall, J. (2023). It ain't near 'bout fair: Re-envisioning the bias and sensitivity review process from a justice-oriented antiracist perspective. *Educational Assessment, doi: 10.1080/10627197.2023.2223924*.

Ray, A. S., & Sinharay, S. (2026). Culturally responsive assessment: Foundations, practices, and pathways toward equity in educational measurement. In S. Sinharay (Ed.), *Encyclopedia of measurement in social sciences*. Elsevier. https://doi.org/10.1016/B978-0-443-26629-4.00217-3

Reckase, M. D. (2009). *Multidimensional item response theory.* Springer.
Richardson, M. W. (1936). The relation between the difficulty and the differential validity of a test. *Psychometrika*, *1*(2), 33–49. https://doi.org/10.1007/bf02288003
Sinharay, S., & Johnson, M. S. (2024). Comparable scores on culturally responsive assessments—how to compute them and when are they accurate? *Journal of Educational Measurement*, *61*(1), 5–46.
Sinharay, S., Johnson, M. S., Bennett, R. E., Lopez, R. M., Sparks, J. R., & Pillarisetti, S. (2025). Investigating elements of culturally responsive assessments in the context of the National Assessment of Educational Progress: An initial exploration. *Educational Measurement: Issues and Practice*, *44*(4), 33–51. https://doi.org/10.1111/emip.70008
Sireci, S. G. (1997). Problems and issues in linking assessments across languages. *Educational Measurement: Issues and Practice*, *16*(1), 12–19.
Sireci, S. G. (2020). Standardization and UNDERSTANDardization in educational assessment. *Educational Measurement: Issues and Practice*, *39*(3), 100–105. https://doi.org/10.1111/emip.12377
Sireci, S. G., & Randall, J. (2021). Evolving notions of fairness in testing in the united states. In M. Bunch & B. Clauser (Eds.), *The history of educational measurement* (pp. 111–135). Routledge. https://doi.org/10.4324/9780367815318-6
Sireci, S. G., Rios, J. A., & Powers, S. (2016). Comparing scores from tests administered in different languages. In L. L. Cook & N. J. Dorans (Eds.), *Fairness in educational assessment and measurement* (pp. 181–202). Routledge. https://doi.org/10.4324/9781315774527-12
Stan Development Team. (2025). RStan: The R interface to Stan [R package version 2.32.7]. https://mc-stan.org/
Tavakol, M., Stewart, C., & Sharpe, C. C. (2024). Ensuring fairness in assessment in health professions education: Rapid analysis tools to detect differential item functioning across groups. *International Journal of Medical Education*, *15*, 80–83. https://doi.org/10.5116/ijme.6694.de69
Taylor, C. S. (2022). *Culturally and socially responsible assessment: Theory, research, and practice*. Teachers College Press.
van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*(2), 181–204. https://doi.org/10.3102/10769986031002181
van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308. https://doi.org/10.1007/s11336-006-1478-z
Veale, J. R., & Foreman, D. I. (1983). Assessing cultural bias using foil response data: Cultural variation. *Journal of Educational Measurement*, *20*(3), 249–258. https://doi.org/10.1111/j.1745-3984.1983.tb00203.x
Walker, M. E., Olivera-Aguilar, M., Lehman, B., Laitusis, C., Guzman-Orth, D., & Gholson, M. (2023). Culturally responsive assessment: Provisional principles. *ETS Research Report 23-11*. https://doi.org/https://doi.org/10.1002/ets2.12374
Wang, C., Zhu, R., & Xu, G. (2023). Using Lasso and adaptive Lasso to identify DIF in multidimensional 2PL models. *Multivariate Behavioral Research*, *58*(2), 387–407. https://doi.org/10.1080/00273171.2021.1985950
Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice*, *36*(4), 52–61. https://doi.org/http://dx.doi.org/10.1111/emip.12165
Wright, D. B. (2019). Speed gaps: Exploring differences in response latencies among groups. *Educational Measurement: Issues and Practice*, *38*(4), 87–98. https://doi.org/http://dx.doi.org/10.1111/emip.12286

Appendix: One Set of Items from the CRA

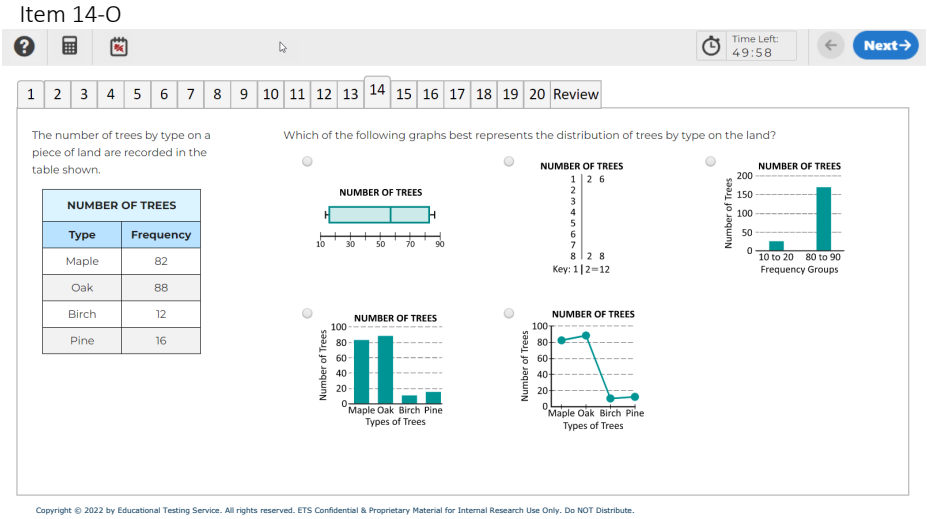

Figure A1. Original Item 14.

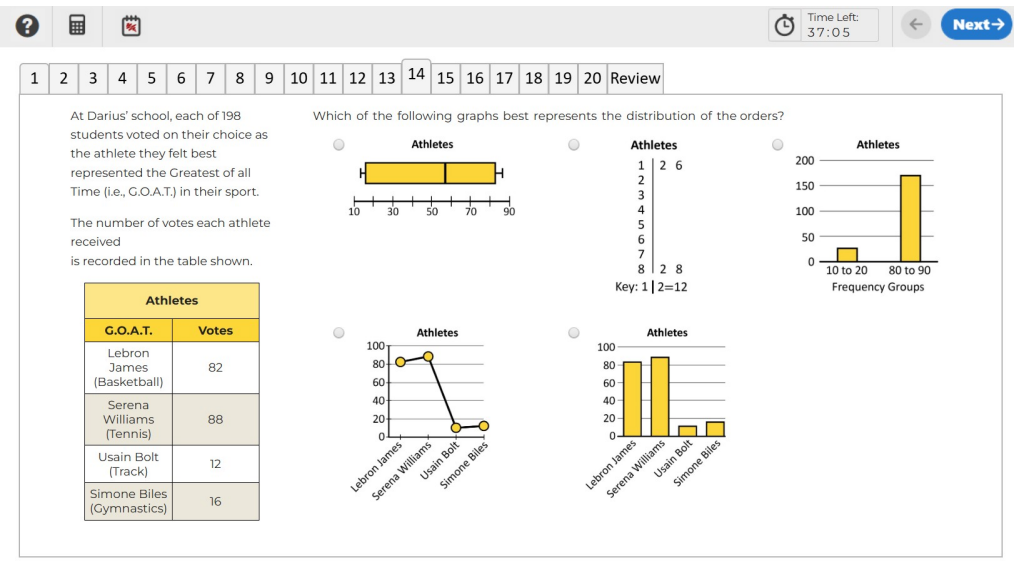

Figure A2. Adapted Version of Item 14.